Critical Values Robust to P-hacking

Adam McCloskey, Pascal Michaillat

May 2023

Abstract. P-hacking is prevalent in reality but absent from classical hypothesis testing theory. As a consequence, significant results are much more common than they are supposed to be when the null hypothesis is in fact true. In this paper, we build a model of hypothesis testing with p-hacking. From the model, we construct critical values such that, if the values are used to determine significance, and if scientists' p-hacking behavior adjusts to the new significance standards, significant results occur with the desired frequency. Such robust critical values allow for p-hacking so they are larger than classical critical values. To illustrate the amount of correction required by p-hacking, we calibrate the model using evidence from the medical sciences. In the calibrated model the robust critical value for any test statistic is the classical critical value for the same test statistic with one fifth of the significance level.

Significance. Scientific journals prefer publishing significant results. Publications, in turn, determine a scientist's career path: promotions, salary, and honors. Scientists therefore have strong incentives to hunt for statistical significance. Such p-hacking reduces the informativeness of hypothesis tests, threatening the credibility of science—leading for instance to the current replication crisis. In this paper, we develop a model of hypothesis testing with p-hacking and use it to construct critical values robust to p-hacking, which guarantee that significant results occur with the desired frequency. As an illustration, we calibrate the model to the medical sciences. For a common two-sided z-test with significance level of 5%, the robust critical value is 2.58—somewhat higher than the classical critical value of 1.96.

McCloskey: University of Colorado–Boulder. Michaillat: Brown University. We thank Isaiah Andrews, Tim Bollerslev, Brian Cadena, Kenneth Chay, Andrew Chen, Garret Christensen, Pedro Dal Bo, Stefano DellaVigna, Peter Hull, Miles Kimball, Jonathan Libgober, Carlos Martins-Filho, Andriy Norets, Emily Oster, Bobak Pakzad-Hurson, Jonathan Roth, Jesse Shapiro, and Yanos Zylberberg for helpful discussions and comments. This work was supported by the Institute for Advanced Study.

1. Introduction

Definition of p-hacking. P-hacking occurs when scientists engage in various behaviors that increase their chances of reporting statistically significant results (Simonsohn, Nelson, and Simmons 2014; Lindsay 2015; Wasserstein and Lazar 2016; Christensen, Freese, and Miguel 2019). Typical p-hacking practices include suppressing inconvenient experiments, halting data collection at a convenient time, dropping inconvenient observations or treatments or outcomes, or choosing convenient statistical specifications.

Prevalence of p-hacking. P-hacking is prevalent in science. Scientists readily admit to it (John, Loewenstein, and Prelec 2012). It is visible in meta-analyses: the distributions of test statistics in entire literatures show that scientists tinker with their analyses to obtain significant results (Hutton and Williamson 2000; Head et al. 2015; Brodeur et al. 2016; Vivalt 2019; Brodeur, Cook, and Heyes 2020; Elliott, Kudrin, and Wuthrich 2022). And it appears by tracking cohorts of scientific studies: studies finding significant results are almost certain to be reported, whereas studies finding insignificant results are likely to remain unreported (Franco, Malhotra, and Simonovits 2014).

Reasons for p-hacking. That p-hacking is so prevalent is unsurprising because scientists face strong incentives to p-hack (Glaeser 2008; Nosek, Spies, and Motyl 2012; Bakker, van Dijk, and Wicherts 2012; Smaldino and McElreath 2016). First, significant results are more rewarded than insignificant ones. This is because scientific journals prefer publishing significant results (Sterling 1959; Bozarth and Roberts 1972; Begg and Berlin 1988; Csada, James, and Espie 1996; Ashenfelter, Harmon, and Oosterbeek 1999; Song et al. 2000; Jennions and Moeller 2002; Ashenfelter and Greenstone 2004; Ioannidis and Trikalinos 2007; Dwan et al. 2008; Ferguson and Brannick 2012; Fanelli, Costas, and Ioannidis 2017; Christensen, Freese, and Miguel 2019; Andrews and Kasy 2019). Publications, in turn, determine a scientist's career path, including promotions, salary, and honorific rewards (Merton 1957; Hagstrom 1965; Skeels and Fairbanks 1968; Katz 1973; Siegfried and White 1973; Tuckman and Leahey 1975; Hansen, Weisbrod, and Strauss 1978; Sauer 1988; Swidler and Goldreyer 1998; Gibson, Anderson, and Tressler 2014; Biagioli and Lippman 2020). Second, scientists enjoy a lot of flexibility in data collection and analysis. Hence, even when the null hypothesis is true, they have ample opportunity to obtain significant results without violating scientific norms (Cole 1957; Armitage 1967; Leamer 1983; Lovell 1983; Simmons, Nelson, and Simonsohn 2011; Humphreys, de la Sierra, and van der Windt 2013; Huntington-Klein et al. 2021).[1]

Problems caused by p-hacking. Despite its prevalence, p-hacking is not accounted for in classical hypothesis testing theory. Therefore, classical critical values set a standard for significance that is too lax: a true null hypothesis is rejected more often than purported by the test's significance level. This is problematic because hypothesis tests are informative only insofar as a true null hypothesis is not rejected more often than the significance level. For instance, hypothesis tests are used to evaluate scientific theories and paradigms (Kuhn 1957; Akerlof and Michaillat 2018). They allow scientists to identify instances when theory does not accord well with empirical observations. Unbridled p-hacking threatens scientific progress. It leads to excessive rejection of established paradigms and to the unwarranted adoption of new paradigms. As such, it threatens the credibility of science. One manifestation of uncontrolled p-hacking is the replication crisis in science (Prinz, Schlange, and Asadullah 2011; Begley and Ellis 2012; Ioannidis et al. 2014; Open Science Collaboration 2015; Camerer et al. 2016; Christensen and Miguel 2018; Benjamin et al. 2018).

Existing corrections for p-hacking. A few corrections for p-hacking in hypothesis testing have been discussed (Anscombe 1954; Lovell 1983; Glaeser 2008). But these corrections take the scientist's p-hacking behavior as fixed, whereas in reality the scientist would change her p-hacking behavior as soon as the correction is implemented. Consider for instance a hypothesis test with a significance level of 5%. Classical critical values are constructed such that if the scientist conducted one experiment, a true null hypothesis would be rejected no more than 5% of the time. But if a scientist conducted more than one experiment, performed hypothesis tests in each experiment separately, and reported the best result, a true null hypothesis would be rejected more often than 5% of the time. Existing corrections take the number of experiments as given and compute a more stringent critical value based on this number. But this is insufficient to resolve the problem. Just as scientists may conduct more than one experiment under the classical critical value, they may conduct more experiments than anticipated under the new critical value, overwhelming the proposed correction.

This paper's correction for p-hacking. In this paper, we start by developing a model of hypothesis testing with p-hacking. We then use the model to construct critical values

[1] Appendix B describes in more detail the literature on p-hacking.

such that, if these values are used to determine significance, and if scientists optimally p-hack in response to the new significance standards, then significant results occur with the desired frequency. Unlike classical critical values, these robust critical values deliver the promised rate of type 1 error. Once the robust critical values are in place, scientists continue to p-hack, but readers can be confident that true null hypotheses are not rejected more often than the advertised significance level.

Model of hypothesis testing with p-hacking. We consider a scientist who tests a hypothesis by conducting an experiment. If she obtains a significant result from the experimental data, she obtains a high payoff. By contrast, if she obtains an insignificant result, she obtains a lower payoff. The difference in payoff between significant and insignificant results reflects the facts that significant results are more likely to be published, and publications yield rewards to scientists. Therefore, if the scientist obtains an insignificant result, and if she still has resources to devote to the project, she has the incentive to conduct another experiment to try to obtain a significant result using the second experiment's data. Conducting a second experiment without revealing the existence of the first experiment constitutes p-hacking.[2]

Optimal p-hacking strategy. Using optimal stopping theory (Ferguson 2007), we find that the scientist's optimal strategy is to conduct experiments until finding a significant result. Not all projects report significant results, however, because the resources that a scientist can devote to any project are finite (Chen 2021). If the scientist runs out of resources before reaching significance, she reports an insignificant result.

Probability of type 1 error. We begin by computing the expected number of experiments run by a scientist when the null hypothesis is true, as a function of the prevailing critical value. From this we compute the probability of type 1 error as a function of the critical value. The critical value influences the rate of type 1 error in two ways. First, it determines the probability that a true null hypothesis is rejected in each experiment—as in classical statistics. Second, it influences the number of experiments that the scientist collects—a feature unique to our model.

[2] Because the number of experiments is not observable, multiple-testing corrections cannot be used to correct for p-hacking.

Computation of robust critical value. From these results we compute the critical value such that type 1 errors occur at the desired rate—given by the significance level. This critical value is robust to p-hacking, and it is given by a nonstandard form of Bonferroni correction: for any test statistic and any significance level, the robust critical value is the classical critical value for the same test statistic with the significance level divided by the expected number of experiments when the robust critical value is in place. Accordingly, the robust critical value is larger than the classical critical value for the same test statistic and significance level. An advantage of the model is that the expected number of experiments when the robust critical value is in place, and the robust critical value itself, are solely determined by two parameters: significance level and probability of completing an experiment before running out of resources.

Numerical illustration. To illustrate the amount of correction that p-hacking might require, we calibrate the completion probability using evidence from the medical sciences (Dwan et al. 2008). We obtain the rule of thumb that the robust critical value for any test statistic is the classical critical value for the same test statistic with one fifth of the significance level. Hence, the robust critical value for a significance level of 5% is the classical critical value for a significance level of 5%/5 = 1%. For a z-test with a significance level of 5%, and similarly for a large-sample t-test with a significance level of 5%, this means that the robust critical value is 2.33 instead of 1.64 if the test is one-sided, and 2.58 instead of 1.96 if the test is two-sided.

2. Model of hypothesis testing with p-hacking

This section develops a simple model of hypothesis testing with p-hacking. A scientist runs experiments with the aim of reaching a significant result. Running experiments takes time, stamina, and money, which are all in finite supply. Because scientists must report results before running out of resources, not all projects yield significant results.

2.1. Hypothesis test

The scientist tests a null hypothesis H_0 against an alternative hypothesis H_1. The data are governed by a different probability distribution under each hypothesis. The scientist sets the test's significance level to $\alpha \in (0, 1)$. The significance level gives the desired probability of type 1 error—the error that occurs when a true null hypothesis is rejected. Common significance levels are 10%, 5%, and 1%.

2.2. Test statistic

To conduct the hypothesis test, the scientist collects a dataset from an experiment. From this dataset she constructs a test statistic T, whose realization is t. Under H_0, the cumulative distribution function of the test statistic is F, its survival function is $S = 1 - F$, and its inverse survival function is $Z = S^{-1}$.[3]

2.3. Classical critical value

The null hypothesis is rejected when the test statistic exceeds the critical value z. If the scientist obtains a test statistic $t > z$, the null hypothesis is rejected: the result is significant. But if she obtains a test statistic $t \leq z$, the null hypothesis cannot be rejected: the result is insignificant. Accordingly, the probability of type 1 error is $S(z)$. The classical critical value is set such that the probability of type 1 error in one single test equals the significance level:

$$S(z) = \alpha, \tag{1}$$

or equivalently $z = Z(\alpha)$.

2.4. Rewards from significant results

The first nonclassical element of the model is the rewards accruing to significant results. To capture the facts that significant results are more likely to be published than insignificant results, and publications yield rewards to scientists, we assume that the expected rewards v^s from a study with significant results are higher than the expected rewards v^i from a study with insignificant results.

2.5. Opportunities for p-hacking

Scientists have ample opportunity to p-hack. However, their resources—time, money, manpower, stamina—are not infinite. Hence, they cannot systematically obtain significant results (Chen 2021). We assume that it takes a random amount of resources to

[3] For simplicity we focus on simple null hypotheses. For composite null hypotheses, we would use the distribution under the null hypothesis's configuration that is the easiest to reject. For example, when testing $H_0 : \mathbb{E}(X) \leq \mu_0$ versus $H_1 : \mathbb{E}(X) > \mu_0$, we would use the distribution of the test statistic at the point $\mathbb{E}(X) = \mu_0$.

conduct an experiment, and the scientist must keep the cumulative resources used below a random limit L. Once the scientist has exhausted more resources than L, she must stop working on the project. The resource limit captures the many resource constraints faced by scientists: limited access to data, limited funding, limited coauthor time, limited time before publication of similar results by competing research teams, limited stamina to work on specific projects, or limited time before the opportunity to work on more promising projects arises. Following Ferguson (2007, p. 4.12), we assume that the resource limit has an exponential distribution with rate $\lambda > 0$, so $\mathbb{P}(L > l) = \exp(-\lambda l)$ for any $l > 0$.

2.6. P-hacking process

Experiments. The experiments are denoted by $n = 0, 1, 2, \ldots, \infty$, with $n = 0$ corresponding to not starting the research project. It takes a random amount of resources to conduct an experiment and collect a dataset. The cumulative amount of resources required to complete $1, 2, \ldots$ experiments is $D_1, D_2, \ldots$ given by a renewal process independent of the resource limit L. That is, the resources required for each experiment, $D_1, D_2 - D_1, D_3 - D_2, \ldots,$ are independent and identically distributed (iid) according to a distribution independent of L.

First experiment. If resources are exhausted before the first experiment is completed, $L < D_1$, the scientist is not able to obtain any results. If the resources are not exhausted when the first experiment is completed, $L > D_1$, the scientist is able to collect a first dataset and construct a test statistic. This first test statistic is T_1, which is independent of the resource variables. The scientist then decides to submit this result to a journal, or to run another experiment.

Nth experiment. If the scientist chooses to run experiment $n \geq 2$, the scientist begins collecting a nth dataset of the same size and drawn from the same underlying population as previous datasets. If resources are exhausted before experiment n is completed, $L < D_n$, the scientist must stop the project before obtaining the nth dataset and submits the best result obtained up to the previous experiment, $\max\{T_1, \ldots, T_{n-1}\}$. If resources are not exhausted, $L > D_n$, the scientist obtains the nth dataset and constructs the nth statistic, T_n, which is iid with $T_1, T_2, \ldots, T_{n-1}$. She may then submit the best of the n test statistics, $\max\{T_1, \ldots, T_n\}$, or she may run yet another experiment.

Infinite p-hacking. $n = \infty$ corresponds to running infinitely many experiments and never reporting any result.

2.7. Completion probability

Following Ferguson (2007, p. 4.13), we introduce the index of the first experiment that cannot be completed before resources are exhausted: $K = \min\{n \geq 1 : D_n > L\}$. Let γ be the probability that the first experiment can be completed:

$$\gamma = \mathbb{P}(D_1 < L) = \mathbb{E}(\exp(-\lambda D_1)).$$

The index K is independent of the test statistics $T_1, T_2, \ldots$, and it has a geometric distribution with success probability $1 - \gamma$, so $\mathbb{P}(K > k) = \gamma^k$ for $k = 0, 1, 2, \ldots$.

2.8. Payoffs

No results. If the scientist does not start the research project, she receives a payoff normalized to $y_0 = 0$. If resources are exhausted before the end of the first experiment, the scientist does not obtain any result, so she receives the same payoff of $y_1 = 0$. If the scientist never concludes the research project and keeps on p-hacking forever, she also receives a payoff $y_\infty = 0$. In all other cases, she receives a positive payoff.

Exhausted resources. The scientist is not able to continue p-hacking once the project resources are exhausted. To capture this constraint, we set to zero all payoffs once resources are exhausted: $y_n = 0$ in any step $n > K$. With these payoffs, the scientist never continues past step K. At step K, the scientist cannot obtain a new test statistic, but she can submit for publication the best test statistic from the previous $K - 1$ hypothesis tests, $\max\{T_1, \ldots, T_{K-1}\}$. If that statistic is significant, the payoff is $y_K = v^s$; if that statistic is not significant, the payoff is $y_K = v^i$.

Non-exhausted resources. Any experiment $n < K$ can be completed before running out of resources, so the scientist can submit the best statistic from the n previous tests, $\max\{T_1, \ldots, T_n\}$. If that statistic is significant, the payoff is $y_n = v^s$; if not, the payoff is $y_n = v^i$.

3. Optimal stopping time

The scientist p-hacks as long as she wishes. At each experiment, she may decide to stop and receive a payoff, or she may decide to continue to the next experiment. If she is able to complete the next experiment, she computes another test statistic. The scientist's problem, which we now solve, is to choose a time to stop p-hacking so as to maximize expected payoffs.

3.1. Scientist's problem

The stopping rule chosen by the scientist, the critical value z, and the random research events determine the random time $N(z)$ at which the scientist stops p-hacking. The problem of the scientist is to choose a stopping time to maximize expected payoffs.

3.2. Reported statistic

As long as she is able to complete at least one hypothesis test, the scientist reports a random statistic $R(z)$ upon stopping. This is the best test statistic that she has been able to obtain through p-hacking. It may be significant or insignificant, and the scientist may be able to publish it or not.

3.3. Characteristics of the optimal stopping time

An optimal stopping time $N(z)$ exists because two conditions are satisfied (Ferguson 2007, chapter 3). Let Y_n denote the random payoff received by the scientist when she stops at time n. First, $Y_n \leq v^s$ a.s., so $\sup_n Y_n < \infty$ a.s. Second, because the resources inevitably run out, $Y_n \overset{as}{\to} 0 = y_\infty$ as $n \to \infty$. Furthermore, the optimal stopping time is given by the principle of optimality of dynamic programming: it is optimal to stop as soon as the payoff is at least as high as the best payoff that can be expected by continuing.

3.4. Finding the optimal stopping time

We find the optimal stopping time by considering the various situations faced by the scientist.

Starting the research project. If the scientist does not start the research project, she receives $Y_0 = 0$. In contrast, if she starts she earns a non-negative payoff: 0 if resources are exhausted before the first experiment is completed; v^i if she obtains an insignificant result; or v^s if she obtains a significant result. Hence it is always optimal to start the research project.

Continuing after insignificant results. How does the scientist behave when she still has resources to allocate to the project? A first possibility is that the result at experiment n and all the results before that are insignificant. Since the best result found by the scientist is insignificant, the scientist earns $Y_n = v^i$ by stopping at experiment n. All possible payoffs are more than the payoff received for an insignificant result, v^i, so all expected payoffs are more than v^i. Since the scientist is expected to obtain more than v^i by continuing, it is not optimal to stop without obtaining a significant result.

Stopping after a significant result. If the result of test n is significant, the best result found by the scientist is significant, so the scientist earns $Y_n = v^s$ by stopping at experiment n. All possible payoffs are less than the payoff received for a significant result, v^s, so all expected payoffs are less than v^s. Hence, the scientist cannot do better by continuing. She optimally stops at experiment n and reports $R(z) = \max\{T_1, \dots, T_n\} > z$. In fact, the principle of optimality indicates that she should stop at the first occurrence of a significant result.

Stopping when resources are depleted. Once resources are depleted, the scientist must stop p-hacking. Hence, she stops at step K if she had not stopped before. There are two possibilities. If $K = 1$, resources are depleted before the first experiment, so the scientist has nothing to report. If $K > 1$, the scientist submits the best test statistic that she has collected. This best result is necessarily insignificant, otherwise she would have stopped before. So she reports $R(z) = \max\{T_1, \dots, T_{K-1}\} \leq z$.

Summary. The optimality principle gives the following results:

LEMMA 1. *The scientist stops when she obtains a significant result or when she runs out of resources, whichever comes first. In the former case the scientist reports a significant result; in the latter case she reports an insignificant result. So there is p-hacking: the scientist never stops at insignificant results, unless she runs out of resources to support the project.*

4. Critical value robust to p-hacking

Based on the scientist's p-hacking strategy, we compute the critical value robust to p-hacking. This critical value ensures that the probability of type 1 error remains below the significance level even as the scientist adjusts her behavior to the critical value itself.

4.1. Distribution of optimal stopping time

We compute the distribution of the optimal stopping time. Since the distribution is used to calculate the critical value, we compute it under the null hypothesis.

Probability of reaching significance at experiment n. Under the null hypothesis, the probability that the test statistic from experiment n reaches the critical value z is given by the test statistic's survival function: $\mathbb{P}(T_n > z) = S(z)$, where $\mathbb{P}$ denotes the probability measure under H_0.

Probability of continuing after experiment n. The scientist continues p-hacking after any experiment if she has not run out of resources during that experiment, which happens with probability γ, and the latest result is insignificant, which happens with probability $1 - S(z)$. The two events are independent, so the probability that the scientist continues p-hacking is $\gamma[1 - S(z)]$. Conversely, the probability that the scientist stops at any experiment is

$$1 - \gamma[1 - S(z)]. \tag{2}$$

Distribution of the stopping time. The probability of stopping at each experiment is constant, given by (2). The optimal stopping time therefore has a geometric distribution with success probability (2). The probability that the optimal stopping time is $n \geq 1$ is

$$\mathbb{P}(N(z) = n) = \left[\gamma - \gamma S(z)\right]^{n-1} \left[1 - \gamma + \gamma S(z)\right].$$

Expected number of experiments. Given that the optimal stopping time has a geometric distribution with success probability (2), we obtain the following result:

PROPOSITION 1. *Under the null hypothesis, the expected number of experiments is*

$$\mathbb{E}(N(z)) = \frac{1}{1 - \gamma[1 - S(z)]}, \tag{3}$$

where $\mathbb{E}$ denotes the expectation operator under H_0. P-hacking is prevalent ($\mathbb{E}(N(z)) > 1$). Scientists p-hack more (higher $\mathbb{E}(N(z))$) when the standards for significance are more stringent (higher z).

Since classical critical values are defined by (1), we infer the following result:

COROLLARY 1. *Under the null hypothesis and with classical critical values, the expected number of experiments is*

$$\mathbb{E}(N(z)) = \frac{1}{1-(1-\alpha)\gamma}. \tag{4}$$

P-hacking is more common (higher $\mathbb{E}(N(z))$) when the significance level is lower (lower α).

P-hacking under the alternative hypothesis. In (4), $1-\alpha$ represents the probability of obtaining an insignificant result from an experiment when the classical critical value is used to determine significance and the null hypothesis is true. When the alternative hypothesis is true instead, the probability of obtaining an insignificant result becomes β, where $1-\beta$ is the power of the hypothesis test. Hence, if the alternative hypothesis is true, the expected number of experiments is $1/(1-\beta\gamma)$. In many fields, hypothesis tests are acceptable only if their power is above 80% (Duflo, Glennerster, and Kremer 2007; Christensen 2018). Setting power to $1-\beta = 80\%$, we find that the expected number of experiments under the alternative is $1/(1-0.2\times\gamma) < 1/(1-0.2) = 1.25$: there is almost no p-hacking. This is unsurprising. If the alternative hypothesis is true and the study is well powered, the null hypothesis is rejected most of the time, which makes p-hacking unnecessary. Hence, if we see a lot of p-hacking, either the alternative hypothesis is false, or the alternative hypothesis is true but tests have low power (Ioannidis 2005).

4.2. Probability of type 1 error

Next, we compute the probability of type 1 error as a function of the critical value.

PROPOSITION 2. *When the critical value is set to z, the probability of finding a type 1 error in a reported hypothesis test is*

$$S^*(z) = \frac{S(z)}{1-\gamma[1-S(z)]}. \tag{5}$$

The probability of type 1 error is larger when scientists p-hack ($S^*(z) > S(z)$). In fact, the *probability of type 1 error grows linearly with the expected number of experiments:*

$$S^*(z) = S(z) \times \mathbb{E}(N(z)). \tag{6}$$

The proof is in appendix A; it relies on appropriate applications of the law of total probability and Bayes' rule. Since classical critical values are defined by (1), we infer the following:

COROLLARY 2. *Under classical critical values, the probability of type 1 error is larger than the significance level:*

$$S^*(z) = \frac{\alpha}{1 - (1 - \alpha)\gamma} > \alpha. \tag{7}$$

When scientists p-hack under classical critical values, the probability of type 1 error exceeds the significance level. Hence, the standard for significance set by classical critical values is too low: significance is reached more often than purported by the test's significance level. This is problematic because hypothesis tests are only informative insofar as true null hypotheses are not rejected more often than the significance level.

4.3. Robust critical value

Effects of critical value on type 1 error rate. Changing the critical value z has two effects on the probability of type 1 error (equation (6)). First, there is a mechanical effect: a higher critical value reduces the probability that a test statistic exceeds it ($S(z)$ is decreasing in z). Second, there is a behavioral effect: the optimal stopping time and reported test statistic are altered by the critical value. When the critical value is larger, scientists p-hack more in hope of reaching significance ($\mathbb{E}(N(z))$ is increasing in z). The behavioral effect was not taken into account by previous corrections for p-hacking (Anscombe 1954; Lovell 1983; Glaeser 2008). The novelty of this analysis is to propose a critical value that accounts for it.

Computing the robust critical value. The robust critical value is such that the probability of type 1 error equals the significance level α when scientists p-hack. Since the probability of type 1 error with p-hacking is given by (5), the robust critical value z^* is implicitly defined by

$$\frac{S(z^*)}{1-\gamma+\gamma S(z^*)}=\alpha. \tag{8}$$

From this definition we obtain the following result (proof details are in appendix A):

PROPOSITION 3. *For any hypothesis test with significance level α, the robust critical value is given by*

$$z^*=Z\left(\alpha\cdot\frac{1-\gamma}{1-\alpha\gamma}\right). \tag{9}$$

The robust critical value is always larger than the classical critical value $Z(\alpha)$.

P-hacking under the robust critical value. The robust critical value corrects the distortion introduced by p-hacking without eliminating p-hacking. In fact, because the significance standards imposed by the robust critical value are more stringent than classical standards, scientists p-hack more under the robust critical value. Combining (3) and (8), we obtain the following corollary:

COROLLARY 3. *The average number of experiments under the robust critical value is*

$$\mathbb{E}(N(z^*))=\frac{1-\alpha\gamma}{1-\gamma}. \tag{10}$$

4.4. Bonferroni correction

Our correction for p-hacking can be formulated as a nonstandard Bonferroni correction:

COROLLARY 4. *The critical value that achieves a significance level α under p-hacking is the critical value that achieves a significance level*

$$\alpha^*=\frac{\alpha}{\mathbb{E}(N(z^*))} \tag{11}$$

under classical conditions.

This relation is obtained by evaluating (6) at z^*, and using $\alpha^*=S(z^*)$ and $S^*(z^*)=\alpha$. Unlike a standard Bonferroni correction, the number of experiments used for the correction is not observed, and it is not the number of experiments prevailing under a standard critical value. Rather, it is the average number of experiments under the robust critical value when the null hypothesis is true. Thanks to the model, we can link this number to the probability γ, which we can calibrate (section 5).

4.5. Influence of the completion probability

Finally, we discuss how the results are influenced by the completion probability γ, which is the main parameter of the model.

Higher completion probability. From equations (3), (5), and (9), we obtain the following:

COROLLARY 5. *Consider a situation with a higher completion probability (higher γ). For a given critical value (z), scientists p-hack more (higher $\mathbb{E}(N(z))$), so type 1 errors are more likely (higher $S^*(z)$). As a result, the robust critical value is higher (higher z^*).*

The corollary indicates that critical values should be higher for research teams with more resources—more time, more money, or more manpower. Research teams with more resources are less likely to be forced to interrupt a study before completion, so they can p-hack more. To control their type 1 error rate properly, a higher critical value is required. The corollary also implies that critical values should be raised when technological progress makes p-hacking easier. An example of such progress is the advent of online surveys and online experiments in social science, which have simplified the task of collecting data. Finally, the corollary implies that critical values should be higher in fields in which p-hacking is easier.

Completion probability of 1. From (4), (7), (9), and (10), we obtain the following results:

COROLLARY 6. *Assume that the completion probability reaches 1 ($\gamma \to 1$). Under the classical critical value, scientists run $1/\alpha$ experiments on average ($\mathbb{E}(N(z)) \to 1/\alpha$), and the probability of type 1 error reaches 1 ($S^*(z) \to 1$). The robust critical value continues to exist but it reaches infinity ($z^* \to \infty$). The average number of experiments under the robust critical value is also infinite ($\mathbb{E}(N(z^*)) \to \infty$).*

The corollary indicates that if scientists can complete any number of experiments, they will continue experimenting until they reach significance. Since all null hypotheses are eventually rejected, the probability of type 1 error is 1. At this limit, scientists successively experiment to reach a foregone conclusion (Anscombe 1954). The robust critical value continues to exist, but it becomes arbitrarily large to offset the arbitrarily large amount of p-hacking.

5. Numerical illustration

To illustrate the amount of correction that p-hacking might require, we calibrate the completion probability γ from the lifecycle of studies in the medical sciences. We then compute the resulting robust critical value.

5.1. Completion probability in the medical sciences

Calibration method. In the model, with probability $1 - \gamma$, the first experiment cannot be completed before running out of resources. The probability $1 - \gamma$ therefore is the share of studies that stop before completion, while the probability γ is the share of studies that are completed. We use data collected by Dwan et al. (2008) to calibrate γ (table 1). Dwan et al. review 16 metastudies that each follow a cohort of medical studies. The studies are followed from protocol approval to publication, so we can measure the fraction of studies that were stopped before completion and thus γ.

Studies that never started. Overall the data include 6903 approved studies. We focus on the 4563 studies whose fate is known—either by surveying the scientists who conducted the studies or by searching the literature. In this pool, 658 were never started, or 658/4563 = 14.4%.

Studies that started but stopped early. In addition, not all the studies that started were completed. Of the 3905 studies that started, 228 were still ongoing when the cohort studies were written, so 3677 studies started and stopped. Of these, 243 stopped early, before any analysis could be conducted. Hence, 243/3677 = 6.6% of the studies that started had to stop before completion.

Calibrated completion probability. Adding the studies that stopped early to those that never started, we find that $14.4\% + (1 - 14.4\%) \times 6.6\% = 20.0\%$ of the approved studies could not be completed. This yields a completion probability of $\gamma = 1 - 20.0\% = 80.0\%$.

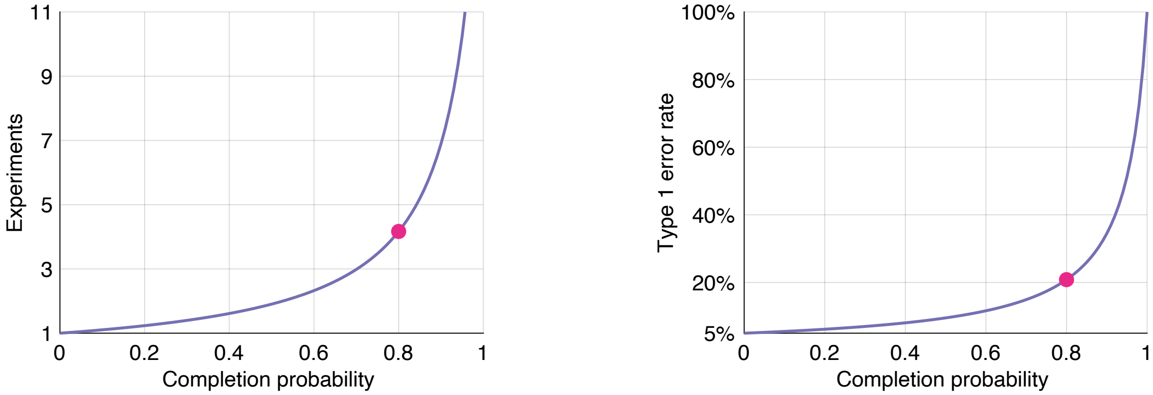
5.2. Obtaining the robust critical value by Bonferroni correction

We now compute the robust critical value using the Bonferroni correction (11) and the completion probability calibrated to the medical sciences, $\gamma = 80\%$.

TABLE 1. Incomplete studies in the medical sciences

		# Studies				
Metastudy	Years	Approved	With information	Never started	Ongoing	Stopped without analysis
Chan et al. (2004a)	1994–2003	304	274	24	2	38
Easterbrook et al. (1991)	1984–1990	715	500	113	42	28
Dickersin, Min, and Meinert (1992)	1980–1988	921	698	184	0	0
Dickersin and Min (1993)	1979–1988	310	270	17	0	0
Stern and Simes (1997)	1979–1992	748	520	100	63	64
Cooper, DeNeve, and Charlton (1997)	1986–	178	159	4	0	2
Wormald et al. (1997)	1963–1997	61	56	5	2	10
Ioannidis (1998)	1986–1996	109	109	0	35	8
Pich et al. (2003)	1997–2001	158	154	11	20	20
Cronin and Sheldon (2004)	1993–1998	101	71	0	0	0
Decullier, Lheritier, and Chapuis (2005)	1994–2002	976	649	68	16	51
Decullier and Chapuis (2006)	1997–2003	142	114	21	29	12
Hahn, Williamson, and Hutton (2002)	1990–1995	56	40	3	0	10
Chan et al. (2004b)	1990–2003	108	105	0	17	0
Ghersi (2006)	1992–	318	318	92	0	0
von Elm et al. (2008)	1988–2006	1698	526	16	2	0
Aggregate	1963–2006	6903	4563	658	228	243

The data for Chan et al. (2004a) appear in figure 3 of Dwan et al. (2008). The data for Easterbrook et al. (1991) appear in figure 4 of Dwan et al. (2008). The data for Dickersin, Min, and Meinert (1992) appear in figure 5 of Dwan et al. (2008). The data for Dickersin and Min (1993) appear in figure 6 of Dwan et al. (2008). The data for Stern and Simes (1997) appear in figure 7 of Dwan et al. (2008). The data for Cooper, DeNeve, and Charlton (1997) appear in figure 8 of Dwan et al. (2008). The data for Wormald et al. (1997) appear in figure 9 of Dwan et al. (2008). The data for Ioannidis (1998) appear in figure 10 of Dwan et al. (2008). The data for Pich et al. (2003) appear in figure 11 of Dwan et al. (2008). The data for Cronin and Sheldon (2004) appear in figure 12 of Dwan et al. (2008). The data for Decullier, Lheritier, and Chapuis (2005) appear in figure 13 of Dwan et al. (2008). The data for Decullier and Chapuis (2006) appear in figure 14 of Dwan et al. (2008). The data for Hahn, Williamson, and Hutton (2002) appear in figure 15 of Dwan et al. (2008). The data for Chan et al. (2004b) appear in figure 16 of Dwan et al. (2008). The data for Ghersi (2006) appear in figure 17 of Dwan et al. (2008). The data for von Elm et al. (2008) appear in figure 18 of Dwan et al. (2008).



A. Prevalence of p-hacking

B. Consequence of p-hacking

FIGURE 1. P-hacking with significance level of 5% and classical critical value

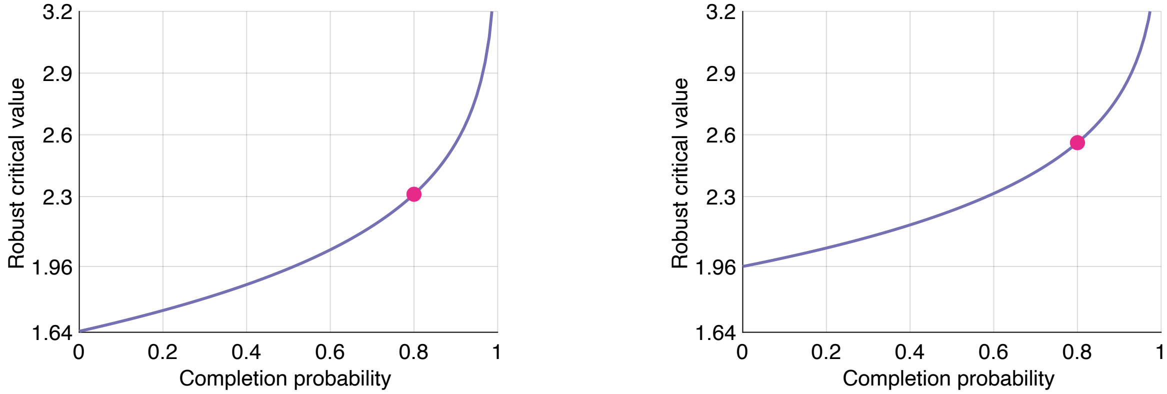
A: The curve gives the expected number of experiments run by a scientist as a function of the probability of completing an experiment when the significance level is 5% and significance is determined by a classical critical value. It is obtained from (4) with $\alpha = 5\%$. B: The curve gives the rate of type 1 error as a function of the probability of completing an experiment when the significance level is 5%, significance is determined by a classical critical value, and the scientist optimally p-hacks. It is obtained from (7) with $\alpha = 5\%$. The pink points indicate the calibrated value of the completion probability: $\gamma = 80\%$.

Formula. Since the significance level α is always less than 10%, and since γ is less than 1, $1 - \alpha\gamma$ is close to 1, and the average number of experiments under the robust critical value is close to $1/(1-\gamma)$ (equation (10)). This gives a simple Bonferroni correction to deal with p-hacking. From (11), we see that the classical significance level α^* required to correct p-hacking is approximately $1-\gamma$ times the desired significance level α:

$$\alpha^* \approx (1-\gamma)\alpha. \tag{12}$$

Numerical application. With $\gamma = 80\%$, (12) implies that the classical significance level required to deal with p-hacking is one fifth of the desired significance level: $\alpha^* = (1-0.8) \times \alpha = \alpha/5$. For instance, the critical value that achieves a significance level of 5% under p-hacking is just the critical value that yields a significance level of 5%/5 = 1% under classical conditions. This rule of thumb works for any test statistic. For a z-test with a significance level of 5%, this means that the robust critical value is 2.33 instead of 1.64 if the test is one-sided, and 2.58 instead of 1.96 if the test is two-sided. These robust critical values also apply to a large sample t-test with a significance level of 5%.

Comparison with the Benjamin et al. (2018) proposal. To address the replication crisis in science, Benjamin et al. (2018) propose that scientists replace the standard signifi-



A. One-sided z-test

B. Two-sided z-test

FIGURE 2. Critical value robust to p-hacking for z-test with significance level of 5%

A: The curve gives the critical value robust to p-hacking for a one-sided z-test with significance level of 5%, as a function of the probability of completing an experiment. It is obtained from (9) where $\alpha = 5\%$ and Z is the inverse survival function for the standard normal distribution. B: The curve gives the critical value robust to p-hacking for a two-sided z-test with significance level of 5%, as a function of the probability of completing an experiment. It is obtained from (9) where $\alpha = 5\%$ and Z is the inverse survival function for the standard half-normal distribution. The pink points indicate the calibrated value of the completion probability: $\gamma = 80\%$.

cance level of 5% by a lower significance level of 0.5%. Such tenfold reduction in the significance level is a more aggressive response to p-hacking than the fivefold reduction obtained in this numerical exercise. However, a tenfold reduction in significance level would be appropriate for a completion probability of $\gamma = 90\%$ (equation (12)). In that way, our analysis provides a theoretical underpinning for proposals to reduce the significance levels used in science. It also links the proposed reductions to the amount of resources available to scientists for p-hacking.

5.3. Additional numerical results

Here we provide additional numerical results. We fix the significance level at 5%.

Prevailing p-hacking. The amount of p-hacking under classical critical values is given by (4). For the completion probability of 80%, the expected number of experiments is 4.2 (figure 1A). Moreover, the amount of p-hacking is increasing with the completion probability. For instance, when the completion probability increases from 70% to 90%, the average number of experiments grows from 3.0 to 6.9.

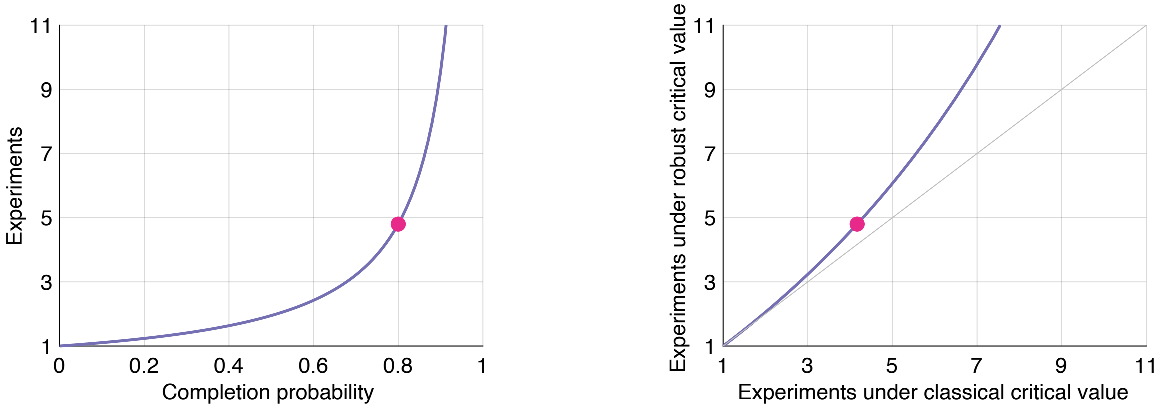
Prevailing probability of type 1 error. The probability of type 1 error under classical critical values is given by (7). For the completion probability of 80%, although the significance level is 5%, the probability of type 1 error is 21% (figure 1B). So in this case, p-hacking quadruples the probability of type 1 error. Moreover, the distortion caused by p-hacking is more severe when the completion probability is larger—because then there is more p-hacking. For instance, when the completion probability increases from 70% to 90%, the probability of type 1 error increases from 15% to 34%.

Robust critical value for one-sided z-test. We calculate the robust critical value when the underlying test statistic has a standard normal distribution under H_0, as in the common z-test, or in a t-test conducted from a large sample. We begin by calculating the robust critical value for a one-sided z-test. The critical value is given by (9) where $\alpha = 5\%$ and Z is the inverse survival function for the standard normal distribution: $Z(x) = \Phi^{-1}(1 - x)$ where Φ is the standard normal cumulative distribution function. For the completion probability $\gamma = 80\%$, the robust critical value is 2.31, almost equal to the value of 2.33 given by the rule of thumb (12) (figure 2A).

Robust critical value for two-sided z-test. Next we calculate the robust critical value for a two-sided z-test. The critical value is now given by (9) where $\alpha = 5\%$ and Z is the inverse survival function for the standard half-normal distribution: $Z(x) = \Phi^{-1}(1 - x/2)$. For the completion probability $\gamma = 80\%$, the robust critical value is 2.56, almost equal to the value of 2.58 given by the rule of thumb (12) (figure 2B).

Sensitivity to the completion probability. Robust critical values are increasing with the completion probability, but they are not very sensitive to it. For instance, as long as the completion probability remains between 70% and 90%, the robust critical value for one-sided z-tests remains between 2.16 and 2.56 (figure 2A), and the robust critical value for two-sided z-tests remains between 2.42 and 2.79 (figure 2B). This is reassuring: robust critical values remain close even in fields with different p-hacking intensity.

P-hacking under robust critical value. The average number of experiments under robust critical value is given by (10). For the completion probability of 80%, the expected number of experiments is 4.8 (figure 3A). Moreover, the amount of p-hacking is increasing with the completion probability. For instance, when the completion probability increases from 70% to 90%, the average number of experiments grows from 3.2 to 9.6.



A. Prevalence of p-hacking

B. Comparison with p-hacking under classical critical value

FIGURE 3. P-hacking with significance level of 5% and robust critical value

A: The curve gives the expected number of experiments run by a scientist as a function of the probability of completing an experiment, when the significance level is 5% and significance is determined by a robust critical value. It is obtained from (10) with $\alpha = 5\%$. B: The curve simultaneously gives the expected number of experiments run by a scientist under classical critical value (horizontal axis) and the expected number of experiments run by a scientist under robust critical value (vertical axis), for any probability of completing an experiment, and for a significance level of 5%. It is obtained from (4) and (10) with $\alpha = 5\%$ and $\gamma \in (0, 1)$. The pink points indicate the calibrated value of the completion probability: $\gamma = 80\%$.

Further, p-hacking is more prevalent under robust critical value than under classical critical value (figure 3B). At the completion probability of 80%, the average number of experiments is 4.2 under classical critical value but 4.8 under robust critical value.

6. Conclusion

We conclude by summarizing our results and discussing several extensions of the model. We also compare our approach to p-hacking with the registration of pre-analysis plans.

6.1. Summary

We develop a simple model of hypothesis testing with p-hacking. We then use the model to construct critical values that correct the excessive rate of type 1 error caused by p-hacking. Once such robust critical values are in place, researchers continue to p-hack, but readers can be confident that true null hypotheses are not rejected more often than the advertised significance level. Robust critical values are larger than classical critical values. As an illustration, we calibrate the p-hacking process with evidence from the medical sciences. We find that the robust critical value for any test and any significance level is the classical critical value for the same test with roughly one fifth of the significance level.

6.2. Extensions of the model

The model presented in this paper is quite stylized, but it can be extended to describe a wider range of p-hacking behaviors. These extensions show that the robust critical values derived in the simple model continue to control false rejections in many situations encountered in practice. For example, the simple model assumes that the scientist constructs test statistics from independent datasets—each obtained from a separate experiment—and therefore obtains independent test statistics. In reality, p-hacking often generates test statistics that are not independent. The robust critical values obtained here remain useful, however, because they maintain the type 1 error rate below the significance level even when p-hacking induces positive dependence of potentially unknown form across test statistics (appendix C). This is the case when scientists pool data across experiments (appendix C.2), when they remove outliers (appendix C.3), when they examine various regression specifications (appendix C.4), or when they examine various instruments (appendix C.5). In addition, the same robust critical values continue to control type 1 error when the model is extended to include a cost of doing research (appendix D) or time discounting (appendix E).

6.3. Comparison with pre-analysis plans

A popular solution to p-hacking is to ask scientists to register pre-analysis plans (Miguel et al. 2014; Christensen and Miguel 2018; Nosek et al. 2018; Adda, Decker, and Ottaviani 2020). Although strict adherence to pre-analysis plans prevents certain forms of p-hacking, it also prevents scientists from engaging in exploratory analysis—a keystone of scientific discovery. By contrast, robust critical values can be used exactly like classical critical values, without preventing exploration of the data. Another concern with pre-analysis plans is that they do not prevent scientists from repeating experiments, as described in the model. A plan could be registered for each experiment until an experiment delivers a significant result, which the scientist would then report with its accompanying pre-analysis plan. Therefore, even when pre-analysis plans are appropriate, it might make sense to use them in conjunction with robust critical values.

References

Adda, Jerome, Christian Decker, and Marco Ottaviani. 2020. "P-hacking in Clinical Trials and How Incentives Shape the Distribution of Results Across Phases." *Proceedings of the National Academy of Sciences* 117 (24): 13386–13392.

Akerlof, George A., and Pascal Michaillat. 2018. "Persistence of False Paradigms in Low-Power Sciences." *Proceedings of the National Academy of Sciences* 115 (52): 13228–13233.

Andrews, Isaiah, and Maximilian Kasy. 2019. "Identification of and Correction for Publication Bias." *American Economic Review* 109 (8): 2766–2794.

Anscombe, Francis J. 1954. "Fixed-Sample-Size Analysis of Sequential Observations." *Biometrics* 10 (1): 89–100.

Armitage, Peter. 1967. "Some Developments in the Theory and Practice of Sequential Medical Trials." In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, edited by Lucien M. Le Cam and Jerzy Neyman, vol. 4, 791–804. Berkeley, CA: University of California Press.

Ashenfelter, Orley C., and Michael Greenstone. 2004. "Estimating the Value of a Statistical Life: The Importance of Omitted Variables and Publication Bias." *American Economic Review* 94 (2): 454–460.

Ashenfelter, Orley, Colm Harmon, and Hessel Oosterbeek. 1999. "A Review of Estimates of the Schooling/Earnings Relationship, with Tests for Publication Bias." *Labour Economics* 6 (4): 453–470.

Bakker, Marjan, Annette van Dijk, and Jelte M. Wicherts. 2012. "The Rules of the Game Called Psychological Science." *Perspectives on Psychological Science* 7 (6): 543–554.

Begg, Colin B., and Jesse A. Berlin. 1988. "Publication Bias: a Problem in Interpreting Medical Data." *Journal of the Royal Statistical Society (Series A)* 151 (3): 419–445.

Begley, C. Glenn, and Lee M. Ellis. 2012. "Raise Standards for Preclinical Cancer Research." *Nature* 483: 531–533.

Benjamin, Daniel J., James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Bjorn Brembs, Lawrence Brown, Colin Camerer et al. 2018. "Redefine Statistical Significance." *Nature Human Behaviour* 2 (1): 6–10.

Biagioli, Mario, and Alexandra Lippman. 2020. "Metrics and the New Ecologies of Academic Misconduct." In *Gaming the Metrics: Misconduct and Manipulation in Academic Research*, edited by Mario Biagioli and Alexandra Lippman, 1–23. Cambridge, MA: MIT Press.

Bozarth, Jerold D., and Ralph R. Roberts. 1972. "Signifying Significant Significance." *American Psychologist* 27 (8): 774–775.

Brodeur, Abel, Nikolai Cook, and Anthony Heyes. 2020. "Methods Matter: P-hacking and Publication Bias in Causal Analysis in Economics." *American Economic Review* 110 (11): 3634–3660.

Brodeur, Abel, Mathias Le, Marc Sangnier, and Yanos Zylberberg. 2016. "Star Wars: the Empirics Strike Back." *American Economic Journal: Applied Economics* 8 (1): 1–32.

Camerer, Colin F., Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jurgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, Emma Heikensten, Felix Holzmeister, Taisuke Imai, Siri Isaksson, Gideon Nave, Thomas Pfeiffer, Michael Razen, and

Hang Wu. 2016. “Evaluating Replicability of Laboratory Experiments in Economics.” *Science* 351 (6280): 433–1436.

Chan, An-Wen, Asbjorn Hrobjartsson, Mette T. Haahr, Peter C. Gotzsche, and Douglas G. Altman. 2004a. “Empirical Evidence for Selective Reporting of Outcomes in Randomized Trials: Comparison of Protocols to Published Articles.” *JAMA* 291 (20): 2457–2465.

Chan, An-Wen, Karmela Krleza-Jeric, Isabelle Schmid, and Douglas G. Altman. 2004b. “Outcome Reporting Bias in Randomized Trials Funded by the Canadian Institutes of Health Research.” *Canadian Medical Association Journal* 171 (7): 735–740.

Chen, Andrew Y. 2021. “The Limits of P-hacking: Some Thought Experiments.” *Journal of Finance* 76 (5): 2447–2480.

Christensen, Garret. 2018. “Manual of Best Practices in Transparent Social Science Research.” https://github.com/garretchristensen/BestPracticesManual/blob/65b77b1991e9b6d5360d3fc6aa2bb7528bedf7ff/Manual.pdf.

Christensen, Garret, Jeremy Freese, and Edward Miguel. 2019. *Transparent and Reproducible Social Science Research: How to Do Open Science*. Oakland, CA: University of California Press.

Christensen, Garret, and Edward Miguel. 2018. “Transparency, Reproducibility, and the Credibility of Economics Research.” *Journal of Economic Literature* 56 (3): 920–980.

Cole, LaMont C. 1957. “Biological Clock in the Unicorn.” *Science* 125 (3253): 874–876.

Cooper, H., K. DeNeve, and K. Charlton. 1997. “Finding The Missing Science: The Fate of Studies Submitted for Review By a Human Subjects Committee.” *Psychological Methods* 2 (4): 447–452.

Cronin, Eugenia, and Trevor Sheldon. 2004. “Factors Influencing the Publication of Health Research.” *International Journal of Technology Assessment in Health Care* 20 (3): 351–355.

Csada, Ryan D., Paul C. James, and Richard H. M. Espie. 1996. “The ‘File Drawer Problem’ of Non-Significant Results: Does It Apply to Biological Research?” *Oikos* 76 (3): 591–593.

Decullier, Evelyne, and Francois Chapuis. 2006. “Impact of Funding on Biomedical Research: A Retrospective Cohort Study.” *BMC Public Health* 6: 165.

Decullier, Evelyne, Veronique Lheritier, and Francois Chapuis. 2005. “Fate Of Biomedical Research Protocols and Publication Bias in France: Retrospective Cohort Study.” *BMJ* 331: 19.

Dickersin, Kay, and Yuan-I Min. 1993. “NIH Clinical Trials and Publication Bias.” Online Journal of Current Clinical Trials 50.

Dickersin, Kay, Yuan-I Min, and Curtis L. Meinert. 1992. “Factors Influencing Publication of Research Results: Follow-up of Applications Submitted to Two Institutional Review Boards.” *JAMA* 267 (3): 374–378.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2007. “Using Randomization in Development Economics Research: A Toolkit.” In *Handbook of Development Economics*, edited by T. Paul Schultz and John A. Strauss, vol. 4, 3895–3962. Amsterdam: Elsevier.

Dwan, Kerry, Douglas G. Altman, Juan A. Arnaiz, Jill Bloom, An-Wen Chan, Eugenia Cronin, Evelyne Decullier, Philippa J. Easterbrook, Erik Von Elm, Carrol Gamble et al. 2008. “Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias.” *PLoS ONE* 3 (8): e3081.

Easterbrook, P. J., R. Gopalan, J. A. Berlin, and D. R. Matthews. 1991. “Publication Bias in Clinical

Research." *Lancet* 337 (8746): 867–872.

Elliott, Graham, Nikolay Kudrin, and Kaspar Wuthrich. 2022. "Detecting P-hacking." *Econometrica* 90 (2): 887–906.

Fanelli, Daniele, Rodrigo Costas, and John P. A. Ioannidis. 2017. "Meta-Assessment of Bias in Science." *Proceedings of the National Academy of Sciences* 114 (14): 3714–3719.

Ferguson, Christopher J., and Michael T. Brannick. 2012. "Publication Bias in Psychological Science: Prevalence, Methods for Identifying and Controlling, and Implications for the Use of Meta-Analyses." *Psychological Methods* 17 (1): 120–128.

Ferguson, Thomas S. 2007. "Optimal Stopping and Applications." https://web.archive.org/web/20200812154935/https://www.math.ucla.edu/~tom/Stopping/Contents.html.

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. "Publication Bias in the Social Sciences: Unlocking the File Drawer." *Science* 345 (6203): 1502–1505.

Ghersi, Davina. 2006. "Issues in the Design, Conduct and Reporting of Clinical Trials That Impact on the Quality of Decision Making." PhD dissertation, School of Public Health, Faculty of Medicine, University of Sydney.

Gibson, John, David L. Anderson, and John Tressler. 2014. "Which Journal Rankings Best Explain Academic Salaries? Evidence from the University of California." *Economic Inquiry* 52 (4): 1322–1340.

Glaeser, Edward L. 2008. "Researcher Incentives and Empirical Methods." In *The Foundations of Positive and Normative Economics: A Hand Book*, edited by Andrew Caplin and Andrew Schotter, chap. 13. New York: Oxford University Press.

Hagstrom, Warren. 1965. *The Scientific Community*. New York: Basic Books.

Hahn, S., P. R. Williamson, and J. L. Hutton. 2002. "Investigation of Within-Study Selective Reporting in Clinical Research: Follow-up of Applications Submitted to a Local Research Ethics Committee." *Journal of Evaluation in Clinical Practice* 8 (3): 353–359.

Hansen, W. Lee, Burton A. Weisbrod, and Robert P. Strauss. 1978. "Modeling the Earnings and Research Productivity of Academic Economists." *Journal of Political Economy* 86 (4): 729–741.

Head, Megan L., Luke Holman, Rob Lanfear, Andrew T. Kahn, and Michael D. Jennions. 2015. "The Extent and Consequences of P-hacking in Science." *PLoS Biology* 13 (3): e1002106.

Humphreys, Macartan, Raul Sanchez de la Sierra, and Peter van der Windt. 2013. "Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration." *Political Analysis* 21 (1): 1–20.

Huntington-Klein, Nick, Andreu Arenas, Emily Beam, Marco Bertoni, Jeffrey R. Bloem, Pralhad Burli, Naibin Chen, Paul Grieco, Godwin Ekpe, Todd Pugatch, Martin Saavedra, and Yaniv Stopnitzky. 2021. "The Influence of Hidden Researcher Decisions in Applied Microeconomics." *Economic Inquiry* 59 (3): 944–960.

Hutton, J. L., and Paula R. Williamson. 2000. "Bias in Meta-Analysis Due to Outcome Variable Selection Within Studies." *Applied Statistics* 49 (3): 359–370.

Ioannidis, John P. A. 1998. "Effect of the Statistical Significance of Results on the Time to Completion and Publication of Randomized Efficacy Trials." *JAMA* 279 (4): 281–286.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *PLoS Medicine* 2 (8): e124.

Ioannidis, John P. A., Sander Greenland, Mark A. Hlatky, Muin J. Khoury, Malcolm R. Macleod, David Moher, Kenneth F. Schulz, and Robert Tibshirani. 2014. "Increasing Value and Reducing Waste in Research Design, Conduct, and Analysis." *Lancet* 383 (9912): 166–175.

Ioannidis, John P.A., and Thomas A. Trikalinos. 2007. "An Exploratory Test for an Excess of Significant Findings." *Clinical Trials* 4 (3): 245–253.

Jennions, Michael D., and Anders P. Moeller. 2002. "Publication Bias in Ecology and Evolution: An Empirical Assessment Using the 'Trim and Fill' Method." *Biological Reviews* 77 (2): 211–222.

John, Leslie K., George Loewenstein, and Drazen Prelec. 2012. "Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling." *Psychological Science* 23 (5): 524–532.

Katz, David A. 1973. "Faculty Salaries, Promotions, and Productivity at a Large University." *American Economic Review* 63 (3): 469–477.

Kuhn, Thomas S. 1957. *The Copernican Revolution.* Cambridge, MA: Harvard University Press.

Leamer, Edward E. 1983. "Let's Take the Con Out of Econometrics." *American Economic Review* 73 (1): 31–43.

Lindsay, D. Stephen. 2015. "Replication in Psychological Science." *Psychological Science* 26 (12): 1827–1832.

Lovell, Michael C. 1983. "Data Mining." *Review of Economics and Statistics* 65 (1): 1–12.

Merton, Robert K. 1957. "Priorities in Scientific Discovery: A Chapter in the Sociology of Science." *American Sociological Review* 22 (6): 635–659.

Miguel, Edward, Colin Camerer, Katherine Casey, Joshua Cohen, Kevin M. Esterling, Alan Gerber, Rachel Glennerster, Don P. Green, Macartan Humphreys, Guido Imbens et al. 2014. "Promoting Transparency in Social Science Research." *Science* 343 (6166): 30–31.

Nosek, Brian A., Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. 2018. "The Preregistration Revolution." *Proceedings of the National Academy of Sciences* 115 (11): 2600–2606.

Nosek, Brian A., Jeffrey R. Spies, and Matt Motyl. 2012. "Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability." *Perspectives on Psychological Science* 7 (6): 615–631.

Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349 (6251): aac4716.

Pich, Judit, Xavier Carne, Joan-Albert Arnaiz, Begona Gomez, Antoni Trilla, and Juan Rodes. 2003. "Role of a Research Ethics Committee In Follow-Up and Publication of Results." *Lancet* 361 (9362): 1015–1016.

Prinz, Florian, Thomas Schlange, and Khusru Asadullah. 2011. "Believe It or Not: How Much Can We Rely on Published Data on Potential Drug Targets?" *Nature Reviews Drug Discovery* 10: 712.

Sauer, Raymond D. 1988. "Estimates of the Returns to Quality and Coauthorship in Economic Academia." *Journal of Political Economy* 96 (4): 855–866.

Siegfried, John J., and Kenneth J. White. 1973. "Financial Rewards to Research and Teaching: A Case Study of Academic Economists." *American Economic Review* 63 (2): 309–315.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant."

Psychological Science 22 (11): 1359–1366.

Simonsohn, Uri, Leif D. Nelson, and Joseph P. Simmons. 2014. “P-curve: A Key to the File-Drawer.” *Journal of Experimental Psychology: General* 143 (2): 534.

Skeels, Jack W., and Robert P. Fairbanks. 1968. “Publish or Perish: An Analysis of the Mobility of Publishing and Nonpublishing Economists.” *Southern Economic Journal* 35 (1): 17–25.

Smaldino, Paul E., and Richard McElreath. 2016. “The Natural Selection of Bad Science.” *Royal Society Open Science* 3 (9): 160384.

Song, F., A. J. Eastwood, S. Gilbody, L. Duley, and A. J. Sutton. 2000. “Publication and Related Biases: A Review.” *Health Technology Assessment* 4 (10): 1–115.

Sterling, Theodore D. 1959. “Publication Decisions and their Possible Effects on Inferences Drawn from Tests of Significance–Or Vice Versa.” *Journal of the American Statistical Association* 54 (285): 30–34.

Stern, Jerome M., and R. John Simes. 1997. “Publication Bias: Evidence of Delayed Publication In a Cohort Study of Clinical Research Projects.” *BMJ* 315: 640.

Swidler, Steve, and Elizabeth Goldreyer. 1998. “The Value of a Finance Journal Publication.” *Journal of Finance* 53 (1): 351–363.

Tuckman, Howard P., and Jack Leahey. 1975. “What Is an Article Worth?” *Journal of Political Economy* 83 (5): 951–967.

Vivalt, Eva. 2019. “Specification Searching and Significance Inflation Across Time, Methods and Disciplines.” *Oxford Bulletin of Economics and Statistics* 81 (4): 797–816.

von Elm, Erik, Alexandra Rollin, Anette Blumle, Karin Huwiler, Mark Witschi, and Matthias Egger. 2008. “Publication and Non-publication of Clinical Trials: Longitudinal Study of Applications Submitted to a Research Ethics Committee.” *Swiss Medical Weekly* 138: 13–14.

Wasserstein, Ronald L., and Nicole A. Lazar. 2016. “The ASA’s Statement on P-values: Context, Process, and Purpose.” *American Statistician* 70 (2): 129–133.

Wormald, R., J. Bloom, J. Evans, and K. Oldfield. 1997. “Publication Bias in Eye Trials.” In *5th Annual Cochrane Colloquium*, Amsterdam.

Appendix A. Proofs

This appendix provides proofs that are omitted in the main text.

A.1. Proof of proposition 2

We start by computing the probability that the reported test statistic $R(z)$ exceeds a critical value z under the null hypothesis. From the law of total probability:

$$\mathbb{P}(R(z) > z) = \sum_{j \geq 1} \mathbb{P}(R(z) > z \mid N(z) = j)\, \mathbb{P}(N(z) = j), \tag{A1}$$

where, according to Bayes' rule,

$$\text{(A2)} \qquad \mathbb{P}(R(z) > z \mid N(z) = j) = \frac{\mathbb{P}(R(z) > z, N(z) = j \mid N(z) > j - 1)}{\mathbb{P}(N(z) = j \mid N(z) > j - 1)}.$$

Then we compute the conditional probability given by (A2). The fact that $N(z) > j - 1$ means that the project resources have not been exhausted during the first $j - 1$ experiments, and that the $j - 1$ test statistics collected have not been significant. Conditional on $N(z) > j - 1$, three events may happen.

First, with probability $1 - \gamma$, resources are exhausted during experiment j. If $j > 1$, then $N(z) = j$ and the scientist reports an insignificant result: $R(z) \leq z$. If $j = 1$, the scientist does not report any result.

Second, with probability γ, resources are not exhausted during experiment j. This creates two subcases. With probability $\gamma S(z)$, the test statistic T_j obtained from experiment j is significant. Then $N(z) = j$ and $R(z) = T_j > z$. With probability $\gamma[1 - S(z)]$, the test statistic T_j obtained from experiment j is insignificant. In that case, $N(z) > j$.

From this case-by-case description, we see that the probability that the scientist stops at experiment j given that she has already completed $j - 1$ experiments is

$$\text{(A3)} \qquad \mathbb{P}(N(z) = j \mid N(z) > j - 1) = 1 - \gamma + \gamma S(z).$$

And the probability that the scientist obtains a significant result from experiment j given that she has already completed $j - 1$ experiments is

$$\text{(A4)} \qquad \mathbb{P}(R(z) > z, N(z) = j \mid N(z) > j - 1) = \mathbb{P}\left(T_j > z, K > j \mid N(z) > j - 1\right) = \gamma S(z).$$

Combining (A2), (A3), and (A4), we find that the probability of reporting a significant result given that the scientist stops the research project at experiment j is

$$\text{(A5)} \qquad \mathbb{P}(R(z) > z \mid N(z) = j) = \frac{\gamma S(z)}{1 - \gamma + \gamma S(z)}.$$

The probability (A5) is independent of j, which greatly simplifies (A1):

$$\mathbb{P}(R(z) > z) = \frac{\gamma S(z)}{1 - \gamma + \gamma S(z)} \left[\sum_{j \geq 1} \mathbb{P}(N(z) = j)\right] = \frac{\gamma S(z)}{1 - \gamma + \gamma S(z)}.$$

Finally, we compute the probability of reporting a significant result given that any result is reported. This conditional probability is given by

$$\mathbb{P}(R(z) > z \mid L > D_1) = \frac{\mathbb{P}(R(z) > z)}{\mathbb{P}(L > D_1)} = \frac{\gamma S(z)}{1 - \gamma + \gamma S(z)} \cdot \frac{1}{\gamma}.$$

To compute the ratio, we use the fact that with probability γ, resources are not exhausted before the end of the first experiment, so some results will be reported, either significant or insignificant.

Therefore, when the critical value is set to z, the probability of type 1 error in a reported study is

$$S^*(z) = \mathbb{P}(R(z) > z \mid L > D_1) = \frac{S(z)}{1 - \gamma + \gamma S(z)}.$$

A.2. Proof of proposition 3

To compute the robust critical value, we rewrite the definition (8) as

$$S(z^*) = \alpha \cdot \frac{1 - \gamma}{1 - \alpha\gamma}.$$

The inverse of the survival function S is the function Z. Inverting S here, we obtain the explicit expression for the robust critical value:

$$z^* = Z\left(\alpha \cdot \frac{1 - \gamma}{1 - \alpha\gamma}\right). \tag{A6}$$

Equation (A6) indicates that the robust critical value always exists. Since $\alpha \in (0, 1)$ and $\gamma \in (0, 1)$, the ratio $(1 - \gamma)/(1 - \alpha\gamma)$ is in $(0, 1)$. Hence, the argument of the inverse survival function Z in (A6) satisfies

$$0 < \alpha \cdot \frac{1 - \gamma}{1 - \alpha\gamma} < \alpha.$$

Accordingly, the argument is in $(0, 1)$. As the domain of the inverse survival function is $(0, 1)$, the robust critical value exists.

From (A6), we can compare the robust critical value to a classical critical value. A classical critical value is defined by $z = Z(\alpha)$, while the robust critical value is defined by (A6). Since the inverse survival function is strictly decreasing, and since the argument of the inverse survival function in (A6) is strictly less than α, we infer that the robust critical value is strictly larger than the classical critical value: $z^* > z$.

Unsurprisingly, the robust critical value is strictly decreasing in the significance level α. Indeed, the argument of the inverse survival function in (A6) is strictly increasing in the significance level $\alpha \in (0, 1)$. Since the inverse survival function itself is strictly decreasing, we infer that the robust critical value is strictly decreasing in the significance level.

Appendix B. Prevalence of p-hacking, and reasons for it

This appendix develops the argument made in the introduction that p-hacking is prevalent in science. It also discusses the reasons behind p-hacking. The first is that p-hacking is rewarded because statistically significant results have greater payoffs than insignificant ones. The second is that p-hacking is not very costly because scientists have a lot of flexibility in their empirical work.

B.1. Prevalence of p-hacking

P-hacking is prevalent in many sciences.

Survey of scientists. A survey of 5964 psychologists at major US universities conducted by John, Loewenstein, and Prelec (2012, table 1) shows that p-hacking is common. 63% of respondents admit to failing to report all outcomes. 56% admit to deciding whether to collect more data after examining whether the results were significant. 46% admit to selectively reporting studies that “worked”. 38% admit to deciding whether to exclude data after looking at the impact of doing so on the results. 28% admit to failing to report all treatments in a study. And 16% admit to stopping data collection earlier than planned after obtaining the desired results.

Lifecycle of studies. Franco, Malhotra, and Simonovits (2014, table 3) track a cohort of 221 experimental studies in the social sciences, from experimental design to publication, and find evidence of p-hacking. Indeed, 64.6% of the studies reporting insignificant results were never written up, whereas only 4.4% of the studies reporting strongly significant results were not written up. This finding indicates that scientists report results selectively: significant results are almost certain to be reported, whereas insignificant results are likely to remain unreported.

Meta-analyses of published studies. The effects of p-hacking also appear in meta-analyses of published studies (Hutton and Williamson 2000; Head et al. 2015; Brodeur et al. 2016; Vivalt 2019; Brodeur, Cook, and Heyes 2020; Elliott, Kudrin, and Wuthrich 2022). The distributions of test statistics or p-values across studies in a literature show that scientists tinker with their econometric specifications in order to obtain significant results.

B.2. Rewards from significant results

Scientists hunt for significant results because such results are more rewarded than insignificant results. The reason is twofold. First, a study presenting significant results is more likely to be published than one presenting insignificant results. Second, a published study yields higher rewards than an unpublished study.

Publication bias. Indeed, scientific journals prefer publishing significant results. Such publication bias was first identified in psychology journals (Sterling 1959; Bozarth and Roberts 1972; Ferguson and Brannick 2012). It has since been observed across the social sciences (Ashenfelter, Harmon, and Oosterbeek 1999; Ashenfelter and Greenstone 2004; Christensen, Freese, and Miguel 2019), medical sciences (Begg and Berlin 1988; Song et al. 2000; Ioannidis and Trikalinos 2007; Dwan et al. 2008), biological sciences (Csada, James, and Espie 1996; Jennions and Moeller 2002), and many other disciplines (Fanelli, Costas, and Ioannidis 2017). Andrews and Kasy (2019, p. 2767) assess the magnitude of the bias in two literatures: experimental economics and psychology. They find that results significant at the 5% level are 30 times more likely to be published than insignificant results.

Rewards from publication. Publications, in turn, determine a scientist's career path, including promotion (Skeels and Fairbanks 1968) and salary (Katz 1973; Siegfried and White 1973; Tuckman and Leahey 1975; Hansen, Weisbrod, and Strauss 1978; Sauer 1988; Swidler and Goldreyer 1998; Gibson, Anderson, and Tressler 2014). In some countries, scientists are also rewarded with cash bonuses as high as $30,000 for publication in top journals (Biagioli and Lippman 2020, p. 6). Publications yield not only material rewards but also honorific rewards (Hagstrom 1965). One such reward is eponymy, "the practice of affixing the name of the scientist to all or part of what he has found" (Merton 1957). Beyond eponymy are prizes, medals, memberships in academies of sciences, and fellowships in learned societies (Merton 1957).

Rewards from significant results. Accordingly, scientists have an incentive to obtain significant results by p-hacking. Formally, let V be the random variable giving the rewards from a completed study. There are several sources of randomness: the study may not be published at all; or it may be published in one of many possible journals, from the most prestigious to the most obscure; even when it is published in a journal of a given standing, the study's impact may vary. The expected rewards from a study with significant results are

$$v^s = \mathbb{E}(V \mid \text{significant}),$$

and those from a study with insignificant results are

$$v^i = \mathbb{E}(V \mid \text{insignificant}).$$

Using the law of iterated expectations, we find

$$\begin{aligned} v^s = {} & \mathbb{E}(V \mid \text{published \& significant}) \times \mathbb{P}(\text{published} \mid \text{significant}) \\ & + \mathbb{E}(V \mid \text{unpublished \& significant}) \times \mathbb{P}(\text{unpublished} \mid \text{significant}). \end{aligned}$$

We note that $\mathbb{P}(\text{unpublished} \mid \text{significant}) + \mathbb{P}(\text{published} \mid \text{significant}) = 1$, and we assume that conditional on the publication status, the rewards are independent from statistical significance. Then we obtain

$$\begin{aligned} v^s = {} & \left[\mathbb{E}(V \mid \text{published}) - \mathbb{E}(V \mid \text{unpublished})\right] \times \mathbb{P}(\text{published} \mid \text{significant}) \\ & + \mathbb{E}(V \mid \text{unpublished}). \end{aligned}$$

Following the same logic, we find

$$\begin{aligned} v^i = {} & \left[\mathbb{E}(V \mid \text{published}) - \mathbb{E}(V \mid \text{unpublished})\right] \times \mathbb{P}(\text{published} \mid \text{insignificant}) \\ & + \mathbb{E}(V \mid \text{unpublished}). \end{aligned}$$

Accordingly, the expected gain from obtaining a significant result is

$$\begin{aligned} v^s - v^i = {} & \left[\mathbb{P}(\text{published} \mid \text{significant}) - \mathbb{P}(\text{published} \mid \text{insignificant})\right] \\ & \times \left[\mathbb{E}(V \mid \text{published}) - \mathbb{E}(V \mid \text{unpublished})\right]. \end{aligned} \tag{A7}$$

Empirically, significant results are more likely to be published than insignificant ones:

$$\mathbb{P}(\text{published} \mid \text{significant}) > \mathbb{P}(\text{published} \mid \text{insignificant}).$$

Moreover, a published study yields higher rewards than an unpublished one:

$$\mathbb{E}(V \mid \text{published}) > \mathbb{E}(V \mid \text{unpublished}).$$

These facts together with (A7) imply that it is beneficial to obtain a significant result:

$$v^s > v^i.$$

B.3. Opportunities for p-hacking

Scientists have a lot of flexibility in data collection and analysis (Huntington-Klein et al. 2021). This flexibility affords them opportunities to obtain significant results, even when the null hypothesis is true. Indeed scientists have found that it is easy to obtain significant results when the null hypothesis is true, without violating scientific norms in biology (Cole 1957), medical science (Armitage 1967, section 4), economics (Leamer 1983; Lovell 1983), psychology (Simmons, Nelson, and Simonsohn 2011), and political science (Humphreys, de la Sierra, and van der Windt 2013).

Appendix C. Other forms of p-hacking

In the model of section 2, the scientist forms separate test statistics from successive experiments. The test statistics are therefore independent. However, a p-hacker often forms test statistics that are positively dependent—when pooling data across experiments or searching across statistical specifications while using the same dataset. This appendix shows that the robust critical values obtained with independent test statistics continue to control the type 1 error rate with positively dependent statistics.

C.1. General p-hacking process

Critical value robust to p-hacking with positively dependent test statistics. We begin by showing that the robust critical values obtained with independent test statistics maintain the rate of type 1 error below the significance level even when test statistics are positively dependent.

Proposition A1. *Assume that the sequence of test statistics* $T_1, \ldots, T_n$ *is positively dependent:*

$$\mathbb{P}(T_n > z \mid T_1, \ldots, T_{n-1} \leq z) \leq \mathbb{P}(T_n > z) = S(z) \tag{A8}$$

for all $z \geq 0$. *Then the probability of type 1 error under the robust critical value* (9) *does not exceed the significance level.*

Proof. The proof proceeds as the proof of proposition 2, with some adjustments. In particular, note that (A1) and (A2) continue to hold and the probability that resources are exhausted at any step k continues to be $1 - \gamma$. However conditional on $N(z) > j - 1$, the probability the test statistic obtained during step k is significant is now bounded above by $\gamma S(z)$ since

$$\mathbb{P}(T_n > z \mid N(z) > j - 1) = \mathbb{P}(T_n > z \mid T_1, \ldots, T_{n-1} \leq z) \leq S(z). \tag{A9}$$

Therefore, (A3) no longer holds but is replaced by

$$\mathbb{P}(N(z) = j \mid N(z) > j - 1) = 1 - \gamma + \gamma\, \mathbb{P}(T_n > z \mid T_1, \ldots, T_{n-1} \leq z). \tag{A10}$$

Similarly, (A4) no longer holds but is replaced by

$$\mathbb{P}(R(z) > z, N(z) = j \mid N(z) > j - 1) = \gamma\, \mathbb{P}(T_n > z \mid T_1, \ldots, T_{n-1} \leq z). \tag{A11}$$

Since the function $x \mapsto x/(1 - \gamma + x)$ is increasing in $x > 0$ for all $\gamma < 1$, (A9), (A10), (A11) and (A2) imply

$$\mathbb{P}(R(z) > z \mid N(z) = j) \leq \frac{\gamma S(z)}{1 - \gamma + \gamma S(z)}$$

so that (A1) implies

$$\mathbb{P}(R(z) > z) \leq \frac{\gamma S(z)}{1 - \gamma + \gamma S(z)}.$$

Applying (8), we obtain

$$\mathbb{P}\big(R(z^*) > z \mid L > D_1\big) = \frac{\mathbb{P}(R(z^*) > z^*)}{\mathbb{P}(L > D_1)} \leq \frac{\gamma S(z^*)}{1 - \gamma + \gamma S(z^*)} \cdot \frac{1}{\gamma} = \frac{S(z^*)}{1 - \gamma + \gamma S(z^*)} = \alpha,$$

which is just the statement in the proposition. □

Condition ensuring positive dependence of t-statistics. In the common case of sequential t-tests, a simple condition on the covariances between successive t-statistics guarantees that proposition A1 applies:

Proposition A2. *Suppose the sequence of test statistics are distributed as follows under H_0: $(T_1, \ldots, T_n) \sim \mathcal{N}(0, \Omega(n))$, where all the variances $\Omega_{1,1}(n)$, ..., $\Omega_{n,n}(n)$ equal 1 and all covariances $\Omega_{1,n}(n)$,..., $\Omega_{n-1,n}(n)$ are non-negative. Then condition* (A8) *is satisfied so proposition A1 applies.*

Proof. We show (A8) holds by showing the conditional probability on the left-hand side is less than the unconditional probability on the right-hand side after further conditioning on any realized value of an additional statistic.

Note that the normally distributed random vector

$$A(n) = [T_1, \ldots, T_{n-1}] - [\Omega_{1,n}(n), \ldots, \Omega_{n-1,n}(n)]T_n$$

is independent of T_n since

$$\begin{aligned}\operatorname{cov}(A(n), T_n) &= \operatorname{cov}([T_1, \ldots, T_{n-1}] - [\Omega_{1,n}(n), \ldots, \Omega_{n-1,n}(n)]T_n, T_n) \\ &= \operatorname{cov}([T_1, \ldots, T_{n-1}], T_n) - [\Omega_{1,n}(n), \ldots, \Omega_{n-1,n}(n)] \operatorname{var}(T_n, T_n) \\ &= [\Omega_{1,n}(n), \ldots, \Omega_{n-1,n}(n)] - [\Omega_{1,n}(n), \ldots, \Omega_{n-1,n}(n)] = 0.\end{aligned}$$

Using the vector $A(n)$, we describe the conditioning event in (A8) as follows:

$$\begin{aligned}\{T_1, \ldots, T_{n-1} \leq z\} &= \left\{[\Omega_{1,n}(n), \ldots, \Omega_{n-1,n}(n)]T_n \leq z - A(n)\right\} \\ &= \left\{ T_n \leq \min_{1 \leq j \leq n-1 : \Omega_{j,n}(n) > 0} \frac{z - A_j(n)}{\Omega_{j,n}(n)}, \max_{1 \leq j \leq n-1 : \Omega_{j,n}(n) = 0} A_j(n) \leq z \right\}.\end{aligned}$$

Since $A(n)$ and T_n are independent, the conditional distribution of the nth t-statistic given the conditioning event in (A8) and the realized value of $A(n)$ is a standard normal truncated from above:

$$T_n \mid \left\{T_1, \ldots, T_{n-1} \leq z, A(n) = a\right\} \sim \xi \mid \xi \leq \mathcal{U}(a),$$

where $\xi \sim \mathcal{N}(0, 1)$ and

$$\mathcal{U}(a) = \min_{1 \leq j \leq n-1 : \Omega_{j,n}(n) > 0} \frac{z - a_j}{\Omega_{j,n}(n)}.$$

Using the properties of the truncated normal distribution, we characterize the conditional probability of type 1 error for the nth t-statistic given non-rejection by the previous t-statistics in the sequence and the realized value of $A(n)$ as

$$\mathbb{P}(T_n > z \mid T_1, \ldots, T_{n-1} \leq z, A(n) = a) = \begin{cases} 1 - \frac{\Phi(z)}{\Phi(\mathcal{U}(a))} & \text{if } z \leq \mathcal{U}(a), \\ 0 & \text{if } z > \mathcal{U}(a) \end{cases}$$

for all a, where Φ denotes the cumulative distribution function of a standard normal random variable. Therefore for any values of a and z,

$$\mathbb{P}(T_n > z \mid T_1, \ldots, T_{n-1} \leq z, A(n) = a) \leq 1 - \Phi(z).$$

But for $F_A(\cdot)$ equal to the cumulative distribution function of $A(n)$,

$$\begin{aligned} \mathbb{P}(T_n > z \mid T_1, \ldots, T_{n-1} \leq z) &= \int_{\mathbb{R}^{n-1}} \mathbb{P}(T_n > z \mid T_1, \ldots, T_{n-1} \leq z, A(n) = a) dF_A(a) \\ &\leq 1 - \Phi(z) = \mathbb{P}(T_n > z) \end{aligned}$$

and we obtain the statement of the proposition. □

The intuition for the proofs is simple. The optimal p-hacking strategy described by lemma 1 remains identical. Indeed, the derivation of the optimal stopping time does not rely on the independence of the test statistics, so it remains valid when the test statistics are dependent. The stochastic properties of the optimal stopping time and reported test statistic do change. But under assumption (A8), we can guarantee that the robust critical value given by (9) keeps the type 1 error rate below the significance level.

Forms of p-hacking generating positively dependent t-statistics. The distributional assumption in proposition A2 is satisfied by the large-sample joint distribution of a sequence of positively correlated t-statistics under the null hypothesis. Such positive correlation appears under several common forms of p-hacking. Suppose that the scientist constructs a general estimator of the form

$$\hat{\mu}_n = \frac{\sum_{j=1}^{m_n} X_{nj} W_{nj}}{\sum_{j=1}^{m_n} X_{nj}^2} \tag{A12}$$

at step n, where m_n is equal to the sample size used in step n. In the subsections that follow, we show that several common estimators in applied work take the form of (A12). Under standard moment conditions on two sets of m_n approximately iid data points $(X_{n1}, \dots, X_{nm_n})$ and $(W_{n1}, \dots, W_{nm_n})$, a bivariate central limit theorem implies the following distributional approximation for large m_n:

$$\frac{1}{\sqrt{m_n}} \begin{pmatrix} \sum_{j=1}^{m_n}[X_{nj} W_{nj} - \mathbb{E}(X_n W_n)] \\ \sum_{j=1}^{m_n}[X_{nj}^2 - \mathbb{E}(X_n^2)] \end{pmatrix} \sim \mathcal{N}(0, \Sigma_n)$$

with

$$\Sigma_n = \begin{pmatrix} \mathbb{E}(X_n^2 W_n^2) - \mathbb{E}(X_n W_n)^2 & \mathbb{E}(X_n^3 W_n) - \mathbb{E}(X_n^2)\,\mathbb{E}(X_n W_n) \\ \mathbb{E}(X_n^3 W_n) - \mathbb{E}(X_n^2)\,\mathbb{E}(X_n W_n) & \mathbb{E}(X_n^4) - \mathbb{E}(X_n^2)^2 \end{pmatrix}.$$

In turn, the delta method implies that for large m_n,

$$\sqrt{m_n}(\hat{\mu}_n - \mu_n) \sim \mathcal{N}(0, \sigma_n^2) \tag{A13}$$

with

$$\mu_n = \frac{\mathbb{E}(X_n W_n)}{\mathbb{E}(X_n^2)}$$
$$\sigma_n^2 = \frac{\mathbb{E}(X_n^2 W_n^2)\,\mathbb{E}(X_n^2)^3 - 2\,\mathbb{E}(X_n^3 W_n)\,\mathbb{E}(X_n^2)\,\mathbb{E}(X_n W_n) + \mathbb{E}(X_n^4)\,\mathbb{E}(X_n W_n)^2}{\mathbb{E}(X_n^2)^4}.$$

By using an estimator of the form (A12), (A13) shows that the scientist is implicitly testing the null hypothesis $H_{0,n} : \mu_n = \mu_{0,n}$ at step n, where the estimand μ_n and its hypothesized value $\mu_{0,n}$ may differ across experiments n, depending upon the context. Under standard moment conditions, the scientist can consistently estimate the large-sample variances σ_n^2, by some estimator $\hat{\sigma}_n^2$. This enables the formation of t-statistics with standard normal distributions under $H_{0,n}$ in large samples:

$$T_n = \frac{\sqrt{m_n}(\hat{\mu}_n - \mu_{0,n})}{\hat{\sigma}_n} \sim \mathcal{N}(0, 1).$$

Given $\hat{\sigma}_i^2$ and $\hat{\sigma}_j^2$ are consistent for σ_i^2 and σ_j^2,

$$\mathrm{cov}(T_i, T_j) \approx \frac{\sqrt{m_i m_j}\,\mathrm{cov}(\hat{\mu}_i, \hat{\mu}_j)}{\sigma_i \sigma_j} \geq 0$$

if and only if $\mathrm{cov}(\hat{\mu}_i, \hat{\mu}_j) \geq 0$. Thus, for estimators of the form (A12), the conditions of proposition A2 hold in large samples when the standard normal approximation for each T_i holds jointly with the others and $\mathrm{cov}(\hat{\mu}_i, \hat{\mu}_j) \geq 0$ for each $i, j = 1, \dots, n$. Sections C.2, C.3, C.4, and C.5 provide common examples of estimators for which these conditions typically hold.

C.2. Pooling data

P-hacking process. The scientist studies a mean parameter $\mu = \mathbb{E}(W)$ for some random variable W. The null hypothesis is $H_0 : \mu = \mu_0$, which does not differ across experiments. The alternative hypothesis is $\mu > \mu_0$. At each step the scientist adds data to the previous dataset; the additional data are independent and collected from the same underlying population. In step n the scientist constructs an estimate $\hat{\mu}_n$ of the parameter by taking a mean from the pooled dataset:

$$\hat{\mu}_n = \frac{1}{m_n} \sum_{j=1}^{m_n} W_j, \tag{A14}$$

where m_n is the size of the pooled dataset, and $W_1, \dots, W_{m_n}$ are iid random variables with mean μ. Using the notation in (A12), we have $X_{nj} = 1$ and $W_{nj} = W_j$ for all n and j.

Verifying the conditions of proposition A2. Since the scientist accumulates data at each step, $m_i > m_j$ for all $i > j$. Hence, using (A14) for $i \geq j$, we obtain

$$\mathrm{cov}(\hat{\mu}_i, \hat{\mu}_j) = \frac{1}{m_i m_j} \sum_{r=1}^{m_j} \sum_{k=1}^{m_i} \mathrm{cov}(W_r, W_k) = \frac{\mathrm{var}(W)}{m_i} \geq 0.$$

Here we used the assumption that $W_1, \dots, W_{m_n}$ are iid, so $\mathrm{cov}(W_r, W_k) = 0$ for all $r \neq k$ and $\mathrm{cov}(W_r, W_r) = \mathrm{var}(W)$ for all r. Furthermore, any finite set of $\hat{\mu}_i$'s have an approximate joint normal distribution in large samples by a standard multivariate central limit theorem. Therefore, the conditions of proposition A2 are satisfied when the scientist p-hacks by pooling data.

C.3. Removing outliers

P-hacking process. The scientist successively removes outliers from a given dataset of size m. At step n, the scientist discards all data points further away than some value c_n from some value χ. She discards more data points at each step so that $c_n < c_q$ for $n > q$. In this scenario, at step n the scientist constructs an estimate $\hat{\mu}_n$ of the parameter by taking a mean from the trimmed sample:

$$\hat{\mu}_n = \frac{\sum_{j=1}^m W_j \, \mathbb{1}(|W_j - \chi| \leq c_n)}{\sum_{j=1}^m \mathbb{1}(|W_j - \chi| \leq c_n)}, \tag{A15}$$

where $\mathbb{1}$ denotes the indicator function, and $W_1, \ldots, W_m$ are iid random variables. The scientist is implicitly testing a different null hypothesis $H_{0,n} : \mu_n = \mu_{0,n}$ at each step n in this example, where

$$\mu_n = \frac{\mathbb{E}(W \, \mathbb{1}(|W - \chi| \leq c_n)}{\mathbb{P}(|W - \chi| \leq c_n)}.$$

Using the notation in (A12), we have $X_{nj} = \mathbb{1}(|W_j - \chi| \leq c_n)$, $W_{nj} = W_j$ and $m_n = m$ for all n and j.

Verifying the conditions of proposition A2. Any finite set of $\sum_{j=1}^m W_j \, \mathbb{1}(|W_j - \chi| \leq c_i)$'s and $\sum_{j=1}^m \mathbb{1}(|W_j - \chi| \leq c_i)$'s have an approximate joint normal distribution in large samples so that the delta method implies the same for any finite set of $\hat{\mu}_i$'s in this example. In addition, the joint normality of the $\hat{\mu}_i$'s and the delta method provide the approximate covariance between any two $\hat{\mu}_i$'s in large samples, as the following proposition shows:

PROPOSITION A3. *For $\hat{\mu}_n$ defined by* (A15) *and a sequence $W_1, W_2, \ldots$ of iid random variables, for any $i \geq j$, $m \operatorname{cov}(\hat{\mu}_i, \hat{\mu}_j)$ converges to*

$$\frac{\operatorname{var}(W \mid |W - \chi| \leq c_i) + \mathbb{E}(W \mid |W - \chi| \leq c_i) \, \mathbb{E}(W \mid |W - \chi| \leq c_j) \, \mathbb{P}(|W - \chi| > c_i) \, \mathbb{P}(|W - \chi| > c_j)}{\mathbb{P}(|W - \chi| \leq c_j)}$$

as $m \to \infty$.

PROOF. A multivariate central limit theorem and delta method imply

$$\begin{aligned} m \operatorname{cov}(\hat{\mu}_i, \hat{\mu}_j) \to & \frac{\operatorname{cov}(W \, \mathbb{1}(|W - \chi| \leq c_i), W \, \mathbb{1}(|W - \chi| \leq c_j))}{\mathbb{P}(|W - \chi| \leq c_i) \, \mathbb{P}(|W - \chi| \leq c_j)} \\ & - \frac{\mathbb{E}(W \, \mathbb{1}(|W - \chi| \leq c_j) \operatorname{cov}(W \, \mathbb{1}(|W - \chi| \leq c_i), \mathbb{1}(|W - \chi| \leq c_j))}{\mathbb{P}(|W - \chi| \leq c_i) \, \mathbb{P}(|W - \chi| \leq c_j)^2} \\ & - \frac{\mathbb{E}(W \, \mathbb{1}(|W - \chi| \leq c_i) \operatorname{cov}(W \, \mathbb{1}(|W - \chi| \leq c_j), \mathbb{1}(|W - \chi| \leq c_i))}{\mathbb{P}(|W - \chi| \leq c_j) \, \mathbb{P}(|W - \chi| \leq c_i)^2} \\ & + \frac{\mathbb{E}(W \, \mathbb{1}(|W - \chi| \leq c_i) \, \mathbb{E}(W \, \mathbb{1}(|W - \chi| \leq c_j) \operatorname{cov}(\mathbb{1}(|W - \chi| \leq c_j), \mathbb{1}(|W - \chi| \leq c_i))}{\mathbb{P}(|W - \chi| \leq c_j)^2 \, \mathbb{P}(|W - \chi| \leq c_i)^2} \end{aligned}$$

as $m \to \infty$. Next we use the definition of covariance, the fact that for $f(w) = w$ or $f(w) = w^2$,

$$\mathbb{E}(f(W) \mid |W - \chi| \leq c_i) = \frac{\mathbb{E}(f(W)\,\mathbb{1}(|W - \chi| \leq c_i))}{\mathbb{P}(|W - \chi| \leq c_i)},$$

and the result that since $c_i < c_j$,

$$\mathbb{1}(|W - \chi| \leq c_i)\,\mathbb{1}(|W - \chi| \leq c_j) = \mathbb{1}(|W - \chi| \leq c_i).$$

From these and standard algebra, we obtain the result of the proposition. □

This proposition shows when the conditions of proposition A2 should hold in large samples. For example, these conditions hold if $\mathbb{E}(W \mid |W - \chi| \leq c_i)$ and $\mathbb{E}(W \mid |W - \chi| \leq c_j)$ have the same sign. It is natural to expect this latter condition to hold in reasonable applications of outlier removal—that is, for reasonable choices of χ and c_n's. For example, suppose that outliers are considered based on deviations from the mean, so $\mathbb{E}(W) = \chi$. Then if W is symmetrically distributed, this condition holds for any choice of c_n since $\mathbb{E}(W \mid |W - \chi| \leq c_n) = \chi$.

C.4. Examining various regression specifications

P-hacking process. The scientist uses ordinary least squares in the standard linear regression model to estimate an effect of interest. A typical effect of interest would be the population value of a regression coefficient. The scientist uses different regression specifications at each p-hacking step, so the parameter of interest differs at each step. Specifically, at step n the scientist uses ordinary least squares to estimate a regression coefficient in a regression of W_n on X_n from two sets of m iid data points $(W_{n1}, \ldots, W_{nm})$ and $(X_{n1}, \ldots, X_{nm})$ so

$$\hat{\mu}_n = \frac{\sum_{j=1}^{m} X_{nj} W_{nj}}{\sum_{j=1}^{m} X_{nj}^2}. \tag{A16}$$

Here, X_n represents the regressor of interest after it has been projected off of the space spanned by the covariates included in the nth regression model, following the procedure described in the Frisch-Waugh-Lovell theorem.

Verifying the conditions of proposition A2. The least squares estimator in (A16) takes the structure of (A12) with $m_n = m$ for all n and therefore satisfies (A13) when, for example,

W_n and X_n have finite fourth moments. In this context, the conditions of proposition A2 therefore hold if $\text{cov}(\hat{\mu}_i, \hat{\mu}_j) \geq 0$ for each $i, j = 1, \ldots, n$, a natural condition for a set of similar regressions. For example, consider two different regressions generating the data

$$W_i = \mu_i X_i + u_i$$
$$W_j = \mu_j X_j + u_j$$

that satisfy standard assumptions such that the least squares estimators of μ_i and μ_j, $\hat{\mu}_i$ and $\hat{\mu}_j$, are jointly asymptotically normally distributed as $m \to \infty$ and centered at μ_i and μ_j with a $\sqrt{m}$ rate of convergence. In this case,

$$m \,\text{cov}(\hat{\mu}_i, \hat{\mu}_j) \to \frac{\mathbb{E}(u_i u_j X_i X_j)}{\mathbb{E}(X_i^2)\,\mathbb{E}(X_j^2)},$$

which is non-negative if and only if $\mathbb{E}(u_i u_j X_i X_j) \geq 0$. This condition naturally holds when the regressors X_i and X_j and regressands W_i and W_j measure similar quantities. In other words, if the scientist estimates similar population regression coefficients at each p-hacking step, the coefficient estimates should be expected to be positively correlated in large samples. This is easiest to see when $\mathbb{E}(u_i u_j | X_j X_j) = \mathbb{E}(u_i u_j)$ (akin to conditional homoskedasticity) since then $\mathbb{E}(u_i u_j X_i X_j) = \text{cov}(u_i, u_j)\,\text{cov}(X_i X_j)$ if an intercept is included in the regression. In this case, $\text{cov}(\hat{\mu}_i, \hat{\mu}_j) \geq 0$ in large samples if both u_i and u_j and X_i and X_j are positively correlated.

The condition $\text{cov}(\hat{\mu}_i, \hat{\mu}_j) \geq 0$ is also testable from observed data: the delta method allows one to compute the approximate covariances between any two $\hat{\mu}_i$'s in large samples for any choices of W_i and X_i. Proposition A3 is an example of such an exercise.

C.5. Examining various instruments

By modifying some of the definitions in the previous example, we can also cover the case in which the scientist uses two-stage least squares to estimate the effect of interest. Assuming that the instruments are both strong and valid, we can modify the definition of X_n to equal the regressor of interest after all regressors have been projected onto the space spanned by the instruments used at the nth p-hacking step, and then the resulting regressor of interest has been projected off of the space spanned by the covariates included in the nth regression model. If the scientist uses the same dependent variable and second stage covariates at each step and only changes the set of instruments used, and if the regression model is correctly specified, the null hypotheses are identical at each step since each μ_n equals the true second stage regression coefficient.

Appendix D. Cost of research

This appendix extends the model of section 2 by introducing a cost of research, incurred with each new experiment. The cost could be monetary or psychological. We find that the robust critical value is not modified by this extension.

D.1. Assumptions

We introduce an expected cost of doing research, c. The cost could be monetary or psychological; it is incurred at each experiment. Because we focus on fields in which research occurs, we assume that c is low enough relative to the rewards from research, v^i and v^s, such that it is optimal for scientists to engage in research.

D.2. Optimal stopping time and robust critical value

Significant result. Since it is optimal to engage in research, the scientist starts a first experiment. With probability γ, the experiment can be completed, and the scientist obtains a test statistic. If the statistic is significant, the scientist obtains v^s, so she stops immediately. Indeed, she cannot obtain a higher payoff by continuing. The same is true in the future too: any time a scientist obtains a significant result, she immediately stops, since it is impossible to obtain a higher payoff later on.

High research cost. What does the scientist decide if the test statistic is insignificant? It depends on the research cost c. If the cost is high enough, the scientist stops right away. This happens when the possibility of obtaining a significant result in the future does not compensate the research cost. In that case, there is no p-hacking: the scientist conducts one experiment and stops, irrespective of the result. The robust critical value is then just the classical critical value.

Low research cost. Since p-hacking is prevalent in reality, the most realistic scenario is that the research cost is low enough so that the scientist runs a new experiment upon obtaining an insignificant result. In that case, because the scientist faces exactly the same situation after each experiment, the scientist continues to p-hack until she obtains a significant result.

Summary. If the research cost is low enough that p-hacking occurs, the presence of the research cost does not modify the scientist's behavior. It is optimal for the scientist to p-hack until she reaches a significant result. Accordingly, everything remains the same in the model—including the robust critical value.

D.3. Computing the cost boundaries

We now compute the expected payoffs from doing research, the cost below which it is optimal to p-hack, and the cost below which it is optimal to engage in research. The expectations of the payoffs depend on the distribution of the test statistic, which in turn depends on which hypothesis is true. We assume that the scientist is conservative and computes the payoff expectations under the null hypothesis.

Continuation value of research. We first compute the continuation value of research for a scientist who has already recorded an insignificant result. We denote this value V^i. Because the scientist's situation is invariant in time, the continuation value is the same at each experiment. When a scientist decides to continue p-hacking, three scenarios are possible. With probability $1 - \gamma$, the scientist cannot complete the experiment and must submit an insignificant result. She then collects v^i. With probability γ, she can complete the experiment. Then with probability $S(z^*)$, her result is significant and she collects v^s. With probability $1 - S(z^*)$, her result is insignificant once again and the continuation value at this point is V^i. In any case, she must incur a cost c to conduct the experiment. Aggregating these scenarios, we obtain the following continuation value:

$$V^i = (1 - \gamma)v^i + \gamma S(z^*)v^s + \gamma[1 - S(z^*)]V^i - c.$$

Hence the continuation value is

$$V^i = \frac{(1 - \gamma)v^i + \gamma S(z^*)v^s - c}{1 - \gamma[1 - S(z^*)]}. \tag{A17}$$

Condition for p-hacking. From the continuation value (A17), we compute the cost below which it is optimal to p-hack. When a scientist has obtained one insignificant result, it is optimal to continue p-hacking if $V^i > v^i$. After a few steps of algebra, this condition becomes

$$c < \gamma S(z^*)(v^s - v^i).$$

Hence, it is optimal to p-hack if the cost of each experiment is below the threshold

$$c^p = \gamma S(z^*)(v^s - v^i).$$

Of course, the cost threshold is higher when significant results are more rewarded relative to insignificant results.

Condition for research. From the continuation value (A17), we also compute the cost below which it is optimal to engage in research. Given that we have normalized the outside option of the scientist to 0, it is optimal to engage in research if the expected value from it is positive.

When a scientist decides to start research, three scenarios are again possible. With probability $1 - \gamma$, the scientist cannot complete the first experiment and cannot submit any result; she then collects 0. With probability γ, she can complete the first experiment. Then with probability $S(z^*)$, her result is significant and she collects v^s. With probability $1 - S(z^*)$, her result is insignificant and the continuation value at this point is V^i. In any case, she must incur a cost c to conduct the experiment.

Aggregating these scenarios, we obtain the initial continuation value:

$$V^r = (1 - \gamma) \times 0 + \gamma S(z^*)v^s + \gamma[1 - S(z^*)]V^i - c.$$

We rewrite the initial continuation value as

$$V^r = \gamma V^i + \gamma S(z^*)(v^s - V^i) - c.$$

Using the value of V^i given by (A17), we finally obtain

$$V^r = \frac{\gamma S(z^*)}{1 - \gamma[1 - S(z^*)]}v^s + \frac{(1 - \gamma)\gamma[1 - S(z^*)]}{1 - \gamma[1 - S(z^*)]}v^i - \frac{1}{1 - \gamma[1 - S(z^*)]} \cdot c.$$

It is optimal to start a research project if $V^r > y_0 = 0$. This condition becomes

$$c < \gamma S(z^*)v^s + (1 - \gamma)\gamma[1 - S(z^*)]v^i.$$

Hence, it is optimal to start research if the cost of each experiment is below the threshold

$$c^r = \gamma S(z^*)v^s + (1-\gamma)\gamma[1-S(z^*)]v^i.$$

The threshold to engage in research is higher than the threshold to engage in p-hacking:

$$c^r = c^p + \gamma[1-\gamma(1-S(z^*))] > c^p.$$

Hence, for all costs between c^p and c^r, scientists engage in research but do not p-hack.

Appendix E. Time discounting

This appendix introduces time discounting into the model of section 2. When the scientist discounts the future, a result submitted early is more valuable than the same result submitted later. Yet, the scientist's behavior and robust critical value are not modified.

E.1. Assumptions

We introduce a discount factor, $\delta \in (0,1)$. The discount factor cost is incurred at each new experiment, so the value of a research result obtained at step n is discounted by δ^n. Because the returns to research are positive without discounting, they also are positive with discounting, so it is optimal for scientists to engage in research.

E.2. Optimal stopping time and robust critical value

Significant result. The scientist immediately stops whenever she obtains a significant result, since it is impossible to obtain a higher payoff in the future.

High discounting. What does the scientist decide if the result is insignificant? It depends on the value of the discount factor δ. If discounting is high enough, the scientist is better off stopping right away. This happens when the possibility of obtaining a significant result in the future does not compensate the time discounting. In that case, there is no p-hacking: the scientist conducts one experiment and stops, irrespective of the result. The robust critical value is then just the classical critical value.

Low discounting. Since p-hacking is prevalent in reality, the most realistic scenario is that discounting is low enough so the scientist starts a new experiment upon obtaining an insignificant result. Then the scientist continues to p-hack until she obtains a significant result, because she faces the same situation after each experiment.

Summary. If time discounting is low enough that p-hacking occurs, the presence of discounting does not modify the scientist's behavior. It is optimal for the scientist to p-hack until she reaches a significant result. Accordingly, everything remains the same in the model—including the robust critical value.

E.3. Computing the discounting boundary

Given that the properties of the model remain the same with discounting, we can use previous results to compute the discount factor below which it is optimal to p-hack.

Continuation value of research. The key step is computing the continuation value of research for a scientist who has already recorded an insignificant result. We denote this value V^i. Because the scientist's situation is invariant in time, this continuation value is the same at each new experiment. When a scientist decides to continue p-hacking, three scenarios are possible. With probability $1-\gamma$, the scientist cannot complete the new experiment and must submit an insignificant result; she then collects δv^i. With probability γ, she can complete the new experiment. Then with probability $S(z^*)$, her result is significant and she collects δv^s; with probability $1-S(z^*)$, her result is insignificant once again and the continuation value at this point is δV^i. Aggregating these scenarios, we obtain the following continuation value:

$$V^i = (1-\gamma)\delta v^i + \gamma S(z^*)\delta v^s + \gamma[1-S(z^*)]\delta V^i.$$

Hence the continuation value is

$$V^i = \delta\frac{(1-\gamma)v^i + \gamma S(z^*)v^s}{1-\delta\gamma[1-S(z^*)]}.$$

Condition for p-hacking. When a scientist has obtained one insignificant result, it is optimal to p-hack if $V^i > v^i$. After a few steps of algebra, this condition becomes

$$\delta > \frac{v^i}{v^i + \gamma S(z^*)(v^s - v^i)}.$$

Hence, it is optimal to p-hack if the discount factor is above the threshold

$$\delta^p = \frac{v^i}{v^i + \gamma S(z^*)(v^s - v^i)}.$$

Of course, the discounting threshold is lower when significant results are more rewarded relative to insignificant results. If insignificant results are not rewarded at all, then scientists p-hack irrespective of discounting.